8-2019

Attention Mechanism for Recognitionin Computer Vision

Alireza Rahimpour
University of Tennessee, arahimpo@vols.utk.edu

Attention Mechanism for Recognition in Computer Vision

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Alireza Rahimpour

August 2019

To my amazing family who has always been there for me...

Acknowledgments

First of all I want to express my sincere gratitude to my genius advisor, Prof. Hairong Qi. I am forever grateful to her not only for all her support, guidance and help throughout my study and all the opportunities that she has provided me, but also for her cheerful, optimistic and patient personality which set her as my role model in life. She showed me how a person can be on top of her carrier while having an amazing personality, help everybody and value the importance of the family. Being away from my family for a long time, she has been the main support for me during my study and there are not enough words to say thank you to her and I really could not have imagined having a better advisor.

Meanwhile, I would like to thank my committee members, Dr. Jens Gregor, Dr. Russell Zaretzki and Dr. Seddik M. Djouadi for their insightful comments on my research. I greatly appreciate their time and input to this dissertation. I also want to thank all my lab-mates for their great help and support, including Wei Wang, Zhibo Wang, Jiajia Luo, Shuangjiang Li, Rui Guo, Liu Liu, Yang Song, Zhifei Zhang, Ali Taalimi, Austin Albright, Ying Qu, Chengcheng Li, Weisheng Tang, Elliot Greenlee, Razieh Kaviani, Fanqi Wang, Ramin Nabati and Taher Naderi. I really value the friendship and all the good memories we built in AICIP years.

Most importantly, my greatest appreciation goes to my mother, Mahboobeh S.H. who is an angel on earth and I owe her everything I have. Thanks to my late father, Dr. Shapour Rahimpour who taught me the importance of education and discipline. His memory will always be with me. Thanks to my wonderful grandparents, Babajoon and Mamanjoon, who have taught me how to work hard and live a fruitful life. Very special thanks to my dear brother, Omid, and my fantastic aunts Mitra, Mahnaz and Marjan for their unconditional love, support and encouragement. I really could not have done it without them.

Abstract

It has been proven that humans do not focus their attention on an entire scene at once when they perform a recognition task. Instead, they pay attention to the most important parts of the scene to extract the most discriminative information. Inspired by this observation, in this dissertation, the importance of attention mechanism in recognition tasks in computer vision is studied by designing novel attention-based models. In specific, four scenarios are investigated that represent the most important aspects of attention mechanism.

First, an attention-based model is designed to reduce the visual features' dimensionality by selectively processing only a small subset of the data. We study this aspect of the attention mechanism in a framework based on object recognition in distributed camera networks. Second, an attention-based image retrieval system (i.e., person re-identification) is proposed which learns to focus on the most discriminative regions of the person's image and process those regions with higher computation power using a deep convolutional neural network. Furthermore, we show how visualizing the attention maps can make deep neural networks more interpretable. Third, a model for estimating the importance of the objects in a scene based on a given task is proposed. More specifically, the proposed model estimates the importance of the road users that a driver (or an autonomous vehicle) should pay attention to in a driving scenario in order to have safe navigation. In this scenario, the attention estimation is the final output of the model. Fourth, an attention-based module and a new loss function in a meta-learning based few-shot learning system is proposed in order to incorporate the context of the task into the feature representations of the samples and increasing the few-shot recognition accuracy.

In this dissertation, we showed that attention can be multi-facet and studied the attention mechanism from the perspectives of feature selection, reducing the computational

cost, interpretable deep learning models, task-driven importance estimation, and context incorporation. Through the study of four scenarios, we further advanced the field of where "attention is all you need".

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Contributions 6

1.3 Dissertation Organization 7

2 Literature Review **8**

2.1 Attention Mechanism 8

2.2 Object Recognition in Distributed Camera Networks 13

2.3 Person Re-identification in Distributed Camera Networks 15

2.4 Object Importance Estimation 20

2.5 Meta-Learning for Few-Shot Learning 21

3 Object Recognition in Distributed Camera Networks **24**

3.1 Introduction 24

3.1.1 Distributed Object Recognition-The Baseline 26

3.2 Compact Representation of the Features 27

3.2.1 Attention-based Feature Selection 27

3.2.2 Generating Low Dimensional Feature Codes 29

3.3 Experiments and Results 30

3.3.1 Datasets 31

3.3.2 Pedestrian Recognition in Surveillance Video 32

3.3.3 Building Recognition in BMW Dataset 33

3.4 Summary 36

4 Attention-based Person Re-identification in Distributed Camera Networks 37
4.1 Introduction . . . 37
4.2 Model Architecture . . . 40
4.2.1 Triplet Loss . . . 40
4.2.2 Gradient-based Attention Network . . . 42
4.3 Experiments and Results . . . 43
4.3.1 Network Design . . . 43
4.3.2 Datasets . . . 44
4.3.3 Evaluation Metric and Results . . . 45
4.3.4 Interpretable Deep Retrieval Model . . . 46
4.4 Conclusion . . . 50

5 Context Aware Road-user importance Estimation (iCARE) 51
5.1 Introduction . . . 51
5.2 Method . . . 54
5.2.1 Important road-user Proposal Generation . . . 54
5.2.2 Context Aware Representation . . . 55
5.3 Experiments . . . 57
5.3.1 Data set . . . 57
5.3.2 Implementation Details . . . 58
5.3.3 Evaluation and Results . . . 59
5.4 Conclusion and Future works . . . 64

6 Class-Discriminative Meta-Learning based Few-Shot Learning 66
6.1 Introduction . . . 66
6.2 Method . . . 71
6.2.1 Few-Shot Classification . . . 71
6.2.2 Structured Support Set Embedding . . . 72
6.2.3 Context-Aware Query Embedding . . . 75
6.2.4 Zero-Shot Learning and Semi-Supervised Adaptation . . . 77
6.3 Experiments and Results . . . 77

6.3.1 Few-Shot Learning . . . 78
6.3.2 Zero-Shot Classification . . . 80
6.3.3 Semi-supervised Adaptation . . . 82
6.3.4 Ablation Study . . . 82
6.4 Conclusion . . . 83

7 Conclusion and Future Works **84**

Bibliography **86**

Appendices **105**

A Publications . . . 106

Vita **108**

List of Tables

3.1 Comparison of recognition accuracy rate of different methods. For all methods compression ratio is set to 2.4 : 1. 35

4.1 Rank1 accuracy (%) comparison of the proposed method to the state-of-the-art. 47

6.1 Number of samples in episodes in different few-shot classification setting for Omniglot dataset during training. 77

6.2 Omniglot few-shot classification. Results are accuracies averaged over 1000 test episodes and with 95% confidence intervals where reported. 79

6.3 miniImageNet few-shot classification. Results are accuracies averaged over 600 test episodes and with 95% confidence intervals where reported. 80

6.4 Zero-shot classification accuracies on CUB-200. 81

6.5 5-way testing accuracy using CDFS method on *mini*Imagenet for the semi-supervised scenario for different number of unlabeled samples per class (n). 82

6.6 Ablation study to evaluate the effect of S3 loss and query encoder in the CDFS model on miniImagenet dataset. 83

List of Figures

1.1 Object recognition in distributed camera networks. The visual features are sent to the base station and the recognition task is performed. 2

1.2 Distributed camera surveillance network illustration of person re-identification. The goal is matching people across non-overlapping camera views at different times (Bedagkar-Gala and Shah, 2014). 3

1.3 Person re-identification challenges. 4

2.1 (a)-The graphical illustration of the model trying to generate the target word y given a source sentence X. (b)-Sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight of the annotation of the j-th source word for the i-th target word, in grayscale (0: black, 1: white) (Bahdanau et al., 2014). 10

2.2 Model learns a words/image alignment for image caption generation in (Xu et al., 2015) 12

2.3 Examples of attending to the correct object (white indicates the attended regions, underlines indicated the corresponding word) (Xu et al., 2015). . . . 12

2.4 Examples of attending to the correct object using the model (Xu et al., 2015). 12

2.5 Distributed object recognition in wireless camera networks. 13

3.1 Distributed Recognition - Baseline 27

3.2 PRID dataset samples 31

3.3 Few sample images of BMW database 32

3.4 Recognition accuracy for different dimensions of the feature histograms using the DFS method . . . 33
3.5 Confusion matrix of pedestrian recognition via DFS on PRID dataset using 340-D features. . . . 34

4.1 The architecture of the proposed Gradient-based Attention Network (GAN) in training phase. . . . 41
4.2 CUHK01 image samples . . . 45
4.3 CUHK03 image samples . . . 45
4.4 Market 1501 image samples . . . 46
4.5 Rank1 retrieval on Market 1501. Top figure shows an example of successful retrieval using our model and the bottom figure shows a fail case for rank1. However in the fail case, GAN can still retrieve the image in rank4. left images are the query and the right images are the ranked images using GAN. . . . 47
4.6 Visualization of the attention map produced by our proposed method . . . 48
4.7 Visualization of the attention map produced by our proposed method (2) . . 49

5.1 Illustration of an ideal road-user importance estimation during a left turn maneuver. In a driving scenario, there can be many road-users. However, when given an ego-vehicle's path, some road-users are more important for decision-making. . . . 52
5.2 The iCARE model exploits local (i.e. appearance, location) of road-users and global (i.e. intention based context) of the scene to estimate importance of respective road-users. . . . 55
5.3 Examples of the images and annotations in our data set . . . 58
5.4 Precision-recall curves (and $F1$ scores) for different experiment settings. Best viewed in color. . . . 60
5.5 Examples of performance comparison of using appearance feature only (yellow) vs iCARE (red) and ground truth (blue). . . . 61

5.6 Comparison of performance of iCARE model (red) vs fusion of appearance, spatial and input future path features (green). The blue bounding boxes show the ground truth annotations for important road-users. The intensity of the red color shows the level of importance of each road-user estimated by iCARE. Best viewed in color. 62
5.7 Ego-vehicle future path prediction error versus distance from ego-vehicle. . . 63
5.8 Three examples of failure cases of iCARE model estimations. (red: iCARE estimation, blue: ground truth, yellow: using only appearance feature for importance estimation.) 64
5.9 Precision-recall curves for iCARE (light blue) and the baseline (red) when trained on the first annotation and test on the second annotation. 65

6.1 Toy example showing the effect of proposed Class-Discriminative Few-Shot learning (CDFS) model on embedding space. 72
6.2 Model architecture for 5-way, 1-shot classification. 73
6.3 Context-aware query embedding architecture. In this example the query embedding $f_\phi(\mathbf{x}_q)$ and the top prototype $\mathbf{c}_1$ in the support set are in class 1. Change of blue color of the query embedding shows how the encoder pulls this feature towards the prototype of class one (i.e., $\mathbf{c}_1$) by incorporating the task context in episodes. In general, during training, the non-linear function g_θ learns how to modify the query embedding based on support set context to achieve optimum classification performance. 76

Chapter 1

Introduction

1.1 Motivation

In this research we study the importance of attention mechanism in recognition tasks in computer vision by designing novel attention-based models. We investigate four different aspects of attention mechanism. The first two aspects are defined in the context of recognition in distributed camera networks since there is an increasing interest in distributed surveillance camera networks due to the growing availability of cheap sensors and processors, and also a growing need for safety and security from the public.

Recognition in distributed camera networks has a wide variety of applications both in public and private environments, such as security, crime prevention, traffic control, accident prediction and detection, and monitoring patients, elderly and children at home. These applications require monitoring indoor and outdoor scenes of airports, train stations, highways, parking lots, stores, shopping malls and offices. Nowadays there are tens of thousands of cameras in a city collecting a huge amount of data on a daily basis. Researchers are urged to develop intelligent systems to efficiently extract information from large scale data.

In chapter 3, we address the problem of transmitting high dimensional visual features in the bandwidth-limited distributed smart camera networks (Figure 1.1). As the concerns about public safety increase in recent years (due to some incidents such as Boston bombing, etc.), researchers have focused more on developing surveillance systems based on distributed

wireless smart cameras. These cameras can cooperate, forming a wireless visual sensing network whose nodes besides visual sensing, also have processing, storage and communication capabilities. Because the smart camera networks have become increasingly more affordable and perform better in balancing the computational power and energy efficiency, they have been employed in many surveillance tasks including distributed object recognition (Redondi et al., 2015), (Christoudias et al., 2008), (Ferrari et al., 2004) and person re-identification (Lisanti et al., 2015), (Ahmed et al., 2015) to name just a few.

However, a major challenge in visual sensor networks is limitation in terms of transmission bandwidth, storage and processing power. In the traditional system design for visual sensor networks, images are acquired and compressed locally at the camera nodes, and then transmitted to the base station which performs the specific analysis tasks (e.g., video surveillance, object recognition, etc.). However, recently, a new paradigm has emerged based on analyze-then-compress, where the visual content is processed locally at the camera nodes, to extract a concise representation constituted by local visual features (e.g., SIFT, SURF, HOG). Such features are then compressed and transmitted to the base station for further analysis. Since the feature-based representation is usually more compact than the pixel-based representation, the analyze-then-compress approach is particularly attractive for those scenarios for which the bandwidth is scarce (Redondi et al., 2015).

Figure 1.1: Object recognition in distributed camera networks. The visual features are sent to the base station and the recognition task is performed.

In this work, we introduce a novel attention-based mechanism to select the most informative visual features in the training set in order to construct a compact representation of the data and then using this compact representation to generate low-dimensional visual features codes. We reduce the features' dimensionality and thus save the network bandwidth in distributed camera networks via the proposed algorithm based on notion of Non-negative Matrix Factorization (NMF) and sparsity.

Furthermore, in chapter 4, we propose a novel framework in order to investigate the effect of attention on the problem of person re-identification (i.e., image retrieval and matching) in distributed camera networks. Recently, person re-identification has gained increasing research interest in the computer vision community due to its importance in multi-camera surveillance systems. In person re-identification, the goal is matching people across non-overlapping camera views at different times. Figure 1.2 shows a typical person re-identification system. The figure shows the top view of a building floor plan and the relative placement of the cameras with respect to the building. Colored dots depict different people

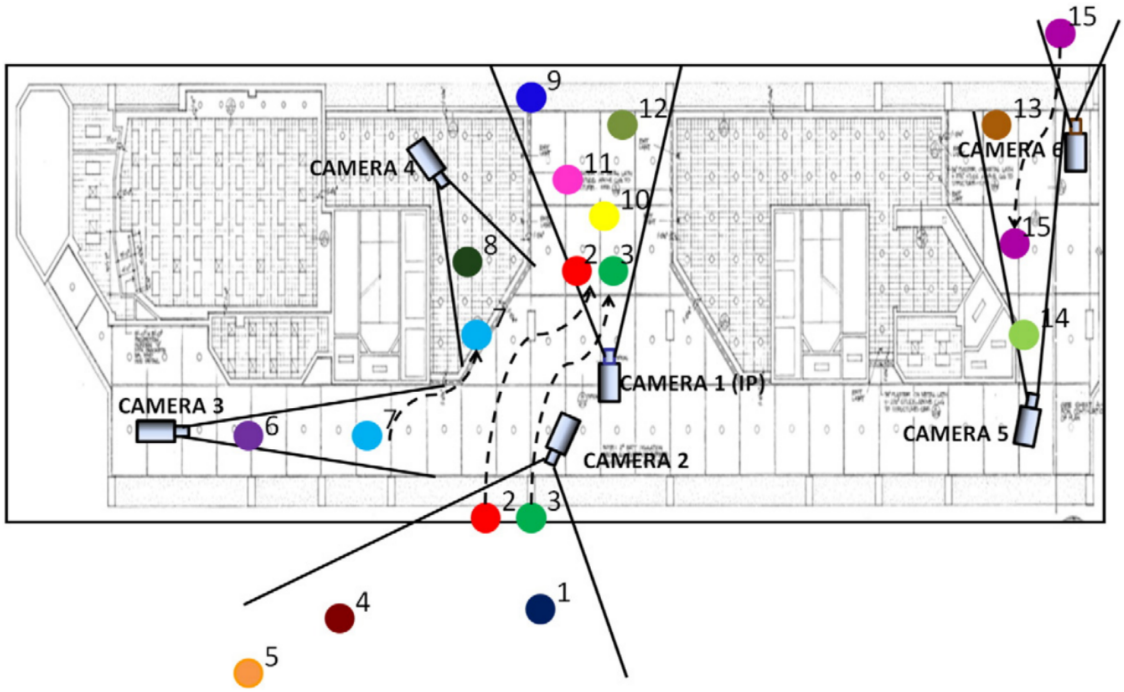


Figure 1.2: Distributed camera surveillance network illustration of person re-identification. The goal is matching people across non-overlapping camera views at different times (Bedagkar-Gala and Shah, 2014).

and numbers besides the dots are the IDs assigned to the people. The dotted lines with arrows represent the directions in which certain people move through the camera network. A typical re-identification system takes as input two images of person's full body, and outputs either a similarity score between the two images or the decision of whether the two images belong to the same identity or not.

Despite all the research efforts, person re-identification remains a challenging problem since a person's appearance can vary significantly when large variations in view angle, human pose, illumination, background clutter and occlusion are involved. In fact, different individuals can share similar appearances and also appearance of the same person can be drastically different in different camera views. Figure 1.3 shows some examples of these challenges.

To address these difficulties, several approaches have been proposed in recent years (e.g., (Cheng et al., 2016a; Varior et al., 2016c; Ahmed et al., 2015; Zhang et al., 2016)) which are mostly based on extracting features from the whole image of the person without focusing on the importance of different parts of the person's image and ignoring the fact that task-relevant information, is often not uniformly distributed across input data.

Human visual attention is well-studied and it is known that human vision is able to focus on a certain region of an image with "high resolution" while perceiving the surrounding image in "low resolution". Inspired by that, we study the impact of attention mechanism in

(a) Cross-view lighting variations (b) Camera viewpoint changes

(c) Clothing similarity (d) Background clutter & occlusions

Figure 1.3: Person re-identification challenges.

solving person re-identification problem. The attention mechanism can significantly reduce the complexity of the person re-identification task, where the network learns to focus on the most informative regions of the scene and ignores the irrelevant parts such as background clutter. Exploiting the attention mechanism in person re-identification task is also beneficial at scaling up the system to large high quality input images. Furthermore, we show how visualizing the attention maps can make the deep neural networks more interpretable. In other words, by visualizing the attention maps we can observe the regions of the input image where the neural network relies on, in order to make a decision. Despite its advantages, exploiting the attention mechanism for person re-identification task have been rarely explored in the literature.

Sometimes the attention estimation is the final output of a recognition model. To study this scenario, in chapter 5, we propose a model for estimating the importance of the objects in a scene based on a given task. More specifically, the proposed model estimates the importance of the road users that a driver (or an autonomous vehicle) should pay attention to in a driving scenario in order to have a safe navigation. Road-users are a critical part of decision-making for both self-driving cars and driver assistance systems. Some road-users, however, are more important for decision making than others because of their respective intentions, ego-vehicles intention and their effects on each other. In this research, we propose a novel approach for road-user importance estimation via fusion of local and global context representations. For local context, we use a hard attention mechanism to consider the appearance of road users (which captures orientation, intention, etc.) and their location relative to ego-vehicle. For global context, we consider the feature map of the last convolutional layer of a model which has been trained to predict the future path of the ego car. Systematic evaluations of our proposed method against several baselines show promising results.

Furthermore, in chapter 6, we propose an attention-based module in a meta-learning based few-shot learning system in order to incorporate the context of the task into the feature representations of the samples in order to increase the few-shot recognition accuracy. Although deep learning-based approaches have been very effective in solving problems with plenty of labeled data, they suffer in tackling problems for which labeled data are scarce. In few-shot classification, the objective is to train a classifier from only a handful of labeled

examples. In this research, we propose an attention-based context-aware query embedding encoder for incorporating support set context into query embedding and generating more discriminative and task-dependent query embeddings. The task-dependent features help the meta learner to learn a distribution over tasks more effectively. Moreover, we propose a few-shot learning framework based on structured margin loss which takes into account the global structure of the support set in order to generate a highly discriminative feature space where the features from distinct classes are well separated in clusters. Extensive experiments based on few-shot, zero-shot and semi-supervised learning on three benchmarks show the advantages of the proposed model compared to state-of-the-art.

1.2 Contributions

In this dissertation novel models are proposed for recognition in computer vision to address the aforementioned aspects of the attention mechanism.

In summary our contributions in this research include:

- Proposing a probabilistic algorithm based on the divergence between the probability distributions of the visual features in order to select the most informative visual features and building a compact and physically meaningful model of the training set in the distributed object recognition framework. We also introduce an NMF-based scheme for calculating the low-dimensional codes for visual feature of each image, before its transmission to the base station and without any communications between the cameras.

- Proposing a CNN-based task-driven attention model which is specifically tailored for the person re-identification task in a triplet architecture. The proposed gradient-based attention model for person re-id is easy to train and the whole network can be trained with back propagation. Moreover, The re-identification network is computationally efficient since it first finds the most discriminative regions in the input image and then performs the deep CNN feature extraction only on these selected regions.

- Designing a new image-based framework for estimating the importance of the road users that a driver (or an autonomous vehicle) should pay attention to in a driving

scenario in order to have a safe navigation. Moreover, we propose a novel context aware architecture and a new way of representing the global context of the scene based on predicting the intention (future path) of the ego-vehicle.

- Designing a novel few-shot learning framework based on meta learning and proposing an attention based context-aware query embedding module which takes into account the support sets context and generates task-dependent feature representations which would help the meta-learner to learn a distribution over tasks more effectively.

- Regularizing the few-shot classification setting with a structured-based margin loss which takes into account the global structure of the support set feature space and learns to explicitly reduce the intra-class variation. This constraint combined with the attention-based encoder, maps the data to a highly discriminative feature space where the few-shot classification is most effective.

- Finally, we quantitatively and qualitatively validate the performance of our proposed models by extensive experiments and comparing them to the state-of-the-art.

1.3 Dissertation Organization

This dissertation is organized as follows: Chapter 2 reviews the works and some basic but essential methods such as Convolutional Neural Networks. Chapter 3 elaborates on attention-based feature encoding in distributed object recognition. Chapter 4 introduces the proposed method for attention-based person re-identification. Chapter 5 describes the novel approach for finding the road-users to which the driver or self-driving car should pay attention. In Chapter 6 a novel attention-based feature encoding for few-shot learning is introduces and chapter 7, concludes the dissertation.

Chapter 2

Literature Review

2.1 Attention Mechanism

Attention Mechanisms in Neural Networks are based on the visual attention mechanism found in humans. Human visual attention is well-studied and while there exist different models, all of them essentially come down to being able to focus on a certain region of an image with "high resolution" while perceiving the surrounding image in "low resolution", and then adjusting the focal point over time. Most traditional computer vision algorithms do not employ attention mechanisms and are indifferent to various parts of the image. With the recent surge of interest in deep neural networks, attention based models have been shown to achieve promising results on several challenging tasks, including caption generation (Xu et al., 2015) and machine translation (Bahdanau et al., 2014) as well as object recognition (Ba et al., 2014). However, most of the attention models proposed so far, require defining an explicit predictive model, whose training can pose challenges due to the non-differentiable cost. Furthermore, some of these models are computationally expensive or need some specific policy algorithms such as reinforcement learning (Ba et al., 2014; Williams, 1992) for training.

Many of these models have employed LSTM based RNNs and have shown good results in learning sequences, but can be computationally expensive. Attention models can be classified into soft attention and hard attention models. Soft attention gives different weights (e.g., using a softmax) to the whole input data based on their importance but hard attention samples an important part of the data. Soft attention models are deterministic and can

be trained using backpropagation, whereas hard attention models are mostly stochastic and can be trained by the reinforcement learning algorithm (Williams, 1992), or by maximizing a variational lower bound or using importance sampling (Ba et al., 2014). Learning hard attention models can become computationally expensive as it requires sampling (however it depends on the type of hard attention). In soft attention approaches, on the other hand, a differentiable mapping can be used from all the locations output to the next input.

We can look at the attention models from saliency aspect. In this way we can categorize the attention to the bottom-up and top-down attention based on how humans focus attention to items present in the environment. The first aspect is called bottom-up processing, also known as stimulus-driven attention or exogenous attention. These describe attentional processing which is driven by the properties of the objects themselves. Some processes, such as motion or a sudden loud noise, can attract our attention in a pre-conscious, or non-volitional way. We attend to them whether we want to or not.

The second aspect is called top-down processing, also known as goal-driven, endogenous attention, attentional control or executive attention. This aspect of our attentional orienting is under the control of the person who is attending. It is mediated primarily by the frontal cortex and basal ganglia (Theeuwes, 1991) as one of the executive functions. Research has shown that it is related to other aspects of the executive functions, such as working memory, and conflict resolution and inhibition (Theeuwes, 1991).

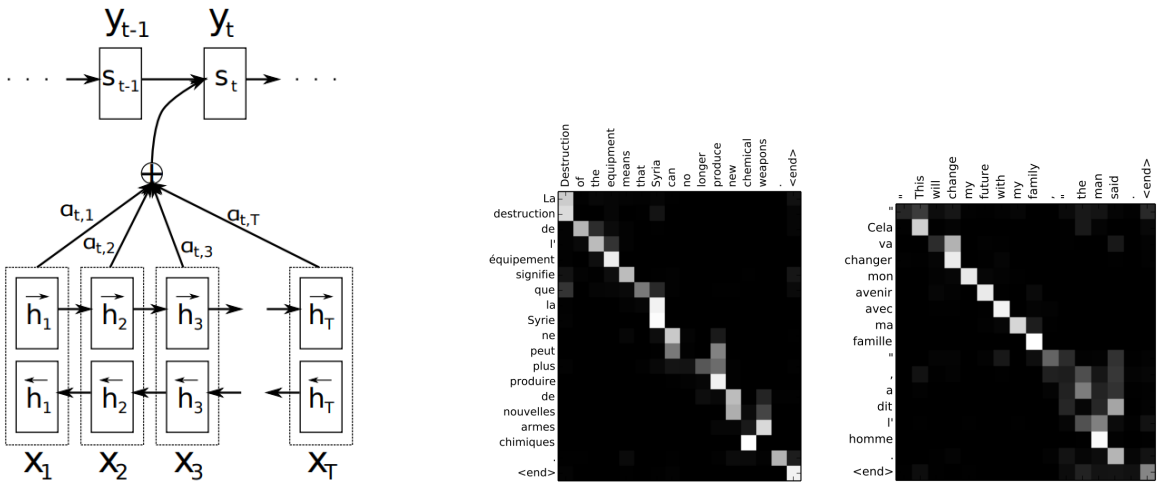
Recently there has been an increasing interest in attention mechanism in different areas. For instance, researchers at Google (Vaswani et al., 2017) proposed the idea of "attention is all you need" (called transformer) as a replacement of RNNs and based on relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for a short time.

Moreover, (Bahdanau et al., 2014) is one of the most important works in using the attention mechanism where they use the attention model for neural machine translation. The models proposed recently for neural machine translation often belong to a family of encoderdecoders and encode a source sentence into a fixed-length vector from which a decoder

generates a translation. In (Bahdanau et al., 2014) they conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoderdecoder architecture, and propose to extend this by allowing a model to automatically (soft attention) search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. Figure 2.1 shows the model and an example of word alignment performed in (Bahdanau et al., 2014).

Self-attention, also known as intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. It has been shown to be very useful in machine reading and image description generation. For example, Cheng et al. (2016b) used self-attention to do machine reading. In machine reading the self-attention mechanism enables us to learn the correlation between the current words and the previous part of the sentence.

(Karpathy et al., 2014) used a multi-resolution CNN architecture to perform action recognition in videos. They mention the concept of fovea but they fix attention to the center of the frame. A recent work of (Xu et al., 2015) used both soft attention and hard



(a) Attention based Model **(b)** Word alignment in translation

Figure 2.1: (a)-The graphical illustration of the model trying to generate the target word y given a source sentence X. (b)-Sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight of the annotation of the j-th source word for the i-th target word, in grayscale (0: black, 1: white) (Bahdanau et al., 2014).

attention mechanisms to generate image descriptions. Their model actually looks at the respective objects when generating their description. Figure 2.2, 2.3 and 2.4 illustrate the model and examples of the image caption generation framework in (Xu et al., 2015).

More recently, (Jaderberg et al., 2015) have proposed a soft attention mechanism called the Spatial Transformer module which they add between the layers of CNNs. Instead of weighting locations using a softmax layer, they apply affine transformations to multiple layers of their CNN to attend to the relevant part and get state-of-the-art results on the Street View House Numbers dataset. (Xu et al., 2015) explored caption generation for image datasets using both soft and hard attention based models and reported state-of-the-art results and most of the video description approaches are based on this work. (Yao et al., 2015) use both 2-D and 3-D CNNs for feature extraction and have a temporal attention mechanism in an LSTM-RNN decoder for generating descriptions of videos. (Yu et al., 2016) use hierarchical RNNs with Gated Recurrent Units (GRUs) and a spatio-temporal attention model (similar to the spatial attention mechanism used by (Xu et al., 2015) to get state-of-the-art results on video description tasks. The hidden state of their GRU-RNN decoder is not conditioned on the weighted video features which gave them higher performance.

In general, it is rather difficult to interpret internal representations learned by deep neural networks. Attention models add a dimension of interpretability by capturing where the model is focusing its attention when performing a particular task. In chapter 4 we propose a method based on using an CNN-based attention model for person re-identification in distributed camera networks which helps the interpretability of the model and also focus on the most discriminative part of the person's image which leads to better recognition accuracy.

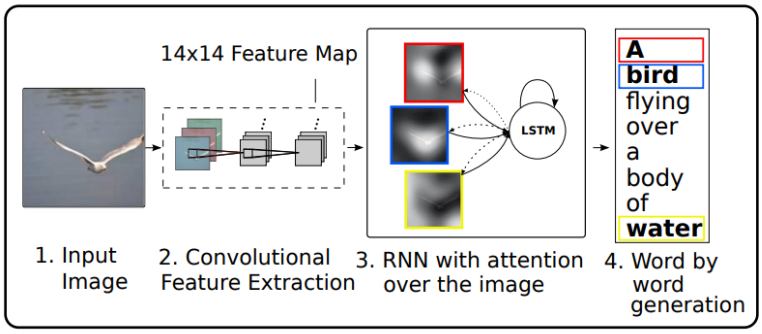


Figure 2.2: Model learns a words/image alignment for image caption generation in (Xu et al., 2015)

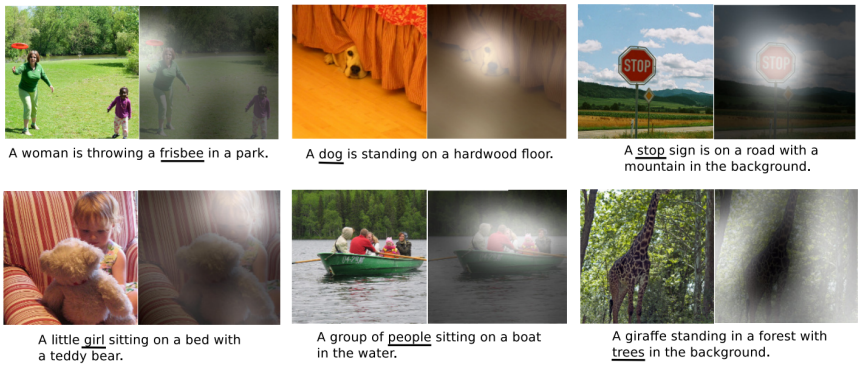


Figure 2.3: Examples of attending to the correct object (white indicates the attended regions, underlines indicated the corresponding word) (Xu et al., 2015).

Figure 2.4: Examples of attending to the correct object using the model (Xu et al., 2015).

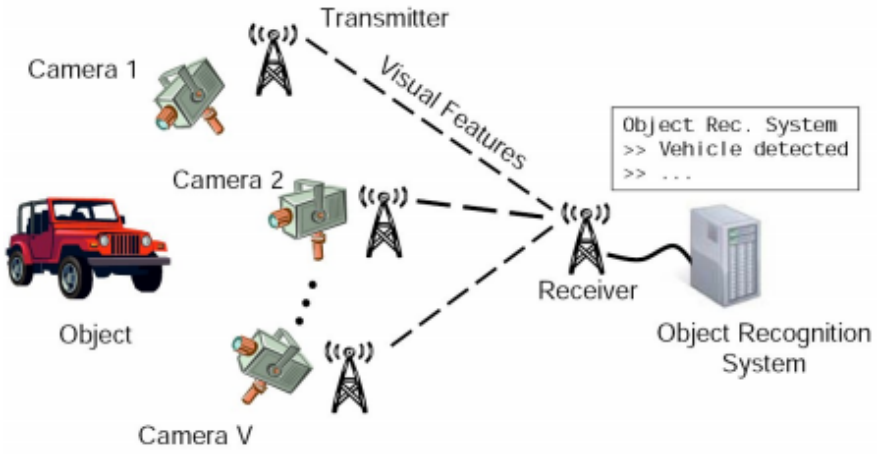


Figure 2.5: Distributed object recognition in wireless camera networks.

2.2 Object Recognition in Distributed Camera Networks

Distributed object recognition in wireless smart camera networks has become focus of interest and numerous of research works have been dedicated to this field especially due to its increasing popularity in surveillance applications (Ye et al., 2013; Rahimpour et al., 2016, 2017c; Taalimi et al., 2017, 2016a). Generally speaking, distributed smart cameras are real-time embedded systems that are able to perform complex computer vision tasks across multiple cameras (Rinner and Wolf, 2008). Figure 2.5 illustrates a distributed object recognition system based on visual features. Traditionally, most computer vision systems have been implemented on workstations, since computer vision applications normally require high-end computational power as well as memory.

Networks of distributed smart cameras can solve computer vision problems in multi-camera applications by providing valuable information through distributed sensing and multi-view processing. Thus, image processing migrates from central workstations to the distributed embedded sensors. The major challenge however, is the limited network bandwidth, making transfer of large amount of visual features infeasible. Therefore, research on distributed object recognition based on visual feature descriptors (e.g, SIFT (Lowe, 1999), SURF (Bay et al., 2006), CHoG (Chandrasekhar et al., 2009)) has been mainly focused on distributed data compression which ensures efficient feature encoding (Christoudias et al.,

2008), (Saxena and Rose, 2010). For instance, (Yang et al., 2010) presented a distributed compression algorithm that encodes the high-dimensional SIFT histograms by Gaussian random projection.

In single view object recognition tasks, when multiple images share common visual features, feature selection algorithms such as (Han and Kim, 2015), (He et al., 2012), (Zhu et al., 2015), and (Qian and Zhai, 2013) are exploited to reduce the redundancy in feature space. However, in distributed object recognition, traditional feature selection approaches are inapplicable as they are computationally expensive and also, the class label is unknown at each camera. Some existing solutions to prune out uninformative features in distributed object recognition, rely on enforcing pairwise epipolar geometry via an expensive structure-from-motion (SfM) procedure (Turcot and Lowe, 2009). In (Naikal et al., 2011a) authors applied Sparse PCA on the feature histograms of each object category in order to select informative features for applications that involve low-quality images from mobile cameras or surveillance camera networks. Moreover, some works such as (Dong et al., 2015; Li et al., 2015) focused on feature engineering and learning (e.g., introducing some new descriptors such as Multi-View HOG). Furthermore, (Yeo et al., 2008) argued that reliable feature correspondence can be established in a much lower dimensional space between cameras, even if the feature vectors are linearly projected onto a random subspace. (Christoudias et al., 2008) studied a SIFT-feature selection algorithm, where the number of SIFT features that need to be transmitted to the base station is reduced by considering the joint distribution of the features among multiple camera views of a common object.

Such solutions are known to break down easily when the camera transformation is large or when the features are extracted from low-quality images. Moreover, most of the existing approaches (e.g., (Turcot and Lowe, 2009), (Christoudias et al., 2008)) require the communication between the smart cameras for selecting the best visual features. In chapter 3 we propose an attention-based feature encoding scheme to address these challenges. Inspired by major success of matrix factorization based models in several fields (Kaviani Baghbaderani and Qi, 2019; Kaviani Baghbaderani et al., 2019; Rahimpour et al., 2015, 2017b; Asadinejad et al., 2018), we define a constrained matrix factorization framework to calculate the low-dimensional feature codes. Nonnegative Matrix Factorization (NMF) (Lee and Seung,

1999) imposes the non-negativity constraint on the factorizing matrices. When all involved matrices are constrained to be nonnegative, NMF allows only additive but not subtractive combinations during the factorization. Such nature can result in parts-based representation of the data, which can discover the hidden components that have specific structures and physical meanings (Lee and Seung, 1999). We will elaborate on this in chapter 3.

2.3 Person Re-identification in Distributed Camera Networks

In general, existing approaches for person re-identification are mainly focused on two aspects: learning a distance metric (Li et al., 2013; Liao et al., 2015; Pedagadi et al., 2013; Su et al., 2015; Xiong et al., 2014) and developing a new feature representation (Varior et al., 2016c; Zhao et al., 2013b; Liao et al., 2010; Ojala et al., 2002; Zhao et al., 2013a; Zheng et al., 2015). However, some interesting recent works based on using Deep Neural Networks tackled the problem of metric learning and new feature representation in a unified framework (Zhang et al., 2016; Cheng et al., 2016a; Ahmed et al., 2015; Wang et al., 2016; Rahimpour et al., 2017a). In the following section we review some of the most important works based on these aspects. In this review we focus on different Re-identification frameworks currently available or likely to be visible in the future, instead of very detailed techniques or architectures.

Metric Learning

In distance metric learning methods, the goal is to learn a metric that emphasizes inter-personal distance and de-emphasizes intra-person distance. The learnt metric is used to make the final decision as to whether a person has been correctly re-identified or not. A comprehensive survey of the metric learning methods can be found in (Yang and Jin, 2006). These metric learning methods are categorized w.r.t supervised learning versus unsupervised learning, global learning versus local learning, etc. In person re-ID, the majority of works fall into the scope of supervised global distance metric learning. The general idea of global metric learning is to keep all the vectors of the same class closer while pushing vectors of different classes further apart. The most commonly used formulation is based on the class of

Mahalanobis distance functions, which generalizes Euclidean distance using linear scalings and rotations of the feature space. In hand-crafted re-ID systems, a good distance metric is critical for its success, because the high-dimensional visual features typically do not capture the invariant factors under sample variances.

Several metric learning algorithms such as Keep It Simple and Straightforward Metric Learning (KISSME) (Köstinger et al., 2012), Locally Adaptive Decision Functions (LADF) (Li et al., 2013), Cross-view Quadratic Discriminant Analysis (XQDA) (Liao et al., 2015), Metric Learning with Accelerated Proximal Gradient (MLAPG) (Su et al., 2015), Local Fisher Discriminant Analysis (LFDA) (Pedagadi et al., 2013) and its kernel variant (k-LFDA) (Xiong et al., 2014) were proposed for person re-identification, achieving remarkable performance in several benchmark datasets.

Features and Representations

In the second group of methods based on developing new feature representation for person re-identification, novel feature representations were proposed to address the challenges such as variations in illumination, pose and view-point (Varior et al., 2016c). The Scale Invariant Local Ternary Patterns (SILTP) (Liao et al., 2010), Local Binary Patterns (LBP) (Ojala et al., 2002), Color Histograms (Zhao et al., 2013a) or Color Names (Zheng et al., 2015) (and the combination of them), are the basis of the majority of these feature representations developed for person re-identification. Compared to the earlier works, handcrafted features have remained more or less the same in recent years. In (Zhao et al., 2014), the LAB color histogram and the SIFT descriptor are extracted from each patch densely sampled with a step size of 5 pixels; this feature is also used in (Shen et al., 2015).

(Das et al., 2014) apply HSV histograms on the head, torso and legs from the silhouette. (Li et al., 2013) also extract local color descriptors from patches but aggregate them using hierarchical Gaussianization to capture spatial information. (Pedagadi et al., 2013) extract color histograms and moments from HSV and YUV spaces before dimension reduction using PCA. (Liu et al., 2014) extract the HSV histogram, gradient histogram and the LBP histogram for each local patch. To improve the robustness of the RGB values against photometric variance, (Yang et al., 2014) introduce the salient color names based color descriptor (SCNCD) for global pedestrian color descriptions. The influence of the background

and different color spaces are also analyzed. In (Liao et al., 2015), Liao et al. propose the local maximal occurrence (LOMO) descriptor, which includes the color and SILTP histograms. Bins in the same horizontal stripe undergo max pooling and a three-scale pyramid model is built before a log transformation.

Apart from directly using low-level color and texture features, another good choice is the attribute-based features which can be viewed as mid-level representations. It is believed that attributes are more robust to image translations compared to low-level descriptors. In (Layne et al., 2012), Layne et al. annotate 15 binary attributes on the VIPeR dataset related to attire and soft biometrics. The low-level color and texture features are used to train the attribute classifiers.

Deep Learning for Re-Identification

In recent years, deep neural networks have been massively used in different recognition, encoding and image synthesis tasks (e.g., (Li et al., 2018; Liu et al., 2017)). Several approaches based on Convolutional Neural Network (CNN) architecture for person re-identification have been proposed and achieved great results (Zhang et al., 2016; Yi et al., 2014; Cheng et al., 2016a; Li et al., 2014; Ahmed et al., 2015; Wang et al., 2016; Xiao et al., 2016; Varior et al., 2016a). In most of the CNN-based approaches for re-identification, the goal is to jointly learn the best feature representation and a distance metric, mostly in a Siamese fashion (Bromley et al., 1993). Siamese networks consist of two identical sub-networks joined at the output which are used for comparing two input images. For learning the network parameters, inputs are therefore given in the form of pairs and the network is optimized by a contrastive loss function (Bromley et al., 1993). The fundamental idea of the contrastive loss function is to attract similar inputs towards each other and repel dissimilar inputs. The first Siamese CNN architecture for person re-identification was proposed in (Yi et al., 2014). By using a Siamese deep neural network, the proposed method in (Yi et al., 2014) can jointly learn the color feature, texture feature and the metric distance. In (Ahmed et al., 2015), Ahmed et al. proposed an improved deep learning architecture which takes pair-wise images as its inputs, and outputs a similarity value indicating whether the two input images depict the same person or not. Their model includes a layer that computes cross-input neighborhood differences to capture local relationships between the two input

images based on their mid-level features, and a patch summary layer to obtain high-level features.

In (Xiao et al., 2016), domain guided dropout was introduced for selecting the appropriate neurons for the images belonging to a given domain. The authors in (Wang et al., 2016) proposed how to model the cross view relationships by jointly learning sub-networks to extract the single image as well as the cross image representations. Local body-part based features and the global features were modeled using a Multi-Channel CNN framework in (Cheng et al., 2016a). Deep Filter Pairing Neural Network (FPNN) was introduced in (Li et al., 2014) to jointly handle misalignment, photometric and geometric transformations, occlusion and cluttered background. Varior, et al. in (Varior et al., 2016a) proposed a matching gate that aimed at comparing features at multiple levels of CNN to boost the local similarities and enhance the discriminative capability of the propagated local features.

Recent advances in Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models (Donahue et al., 2015; Yue-Hei Ng et al., 2015; Sutskever et al., 2014), provide some insights as to how to integrate the contextual information in the model. In (Pei et al., 2016) and (Deng et al., 2009) it has been shown that LSTM cells can detect salient keywords relevant to a topic (i.e., context) from sentences or speech inputs. The extracted salient contextual information can further enhance the discriminative power of the learned local feature representations. In addition to capturing the contextual dependency, LSTM can also selectively allow or block the information flow through the network by using its advanced multiplicative interactions in the cell (Pei et al., 2016), (Yue-Hei Ng et al., 2015). In (Varior et al., 2016b), a Siamese LSTM architecture that can process image regions sequentially and enhance the discriminative capability of local feature representation is presented for person re-identification.

Furthermore, a new trend has been started recently based on the video-based person re-identification and taking advantage of temporal information in the video frames (Liu et al., 2015; Wang et al., 2014; McLaughlin et al., 2016). In practice, video-based person re-identification provides a more natural way for person re-identification, where most often videos are the actual input to the surveillance systems. Furthermore, video-based methods can utilize extra space-time information, which contains much more rich cues about the

identity of the person. In fact, given the availability of sequences of images, temporal priors in relation to person's motion, such as gait and pose are captured which may assist in solving the re-identification problem for difficult cases. In fact, sequences of images provide more samples of a pedestrians appearance, where each sample may contain different pose and viewpoint, and thus allows a more reliable appearance-based model to be constructed. However, making use of time series also brings about new challenges to re-identification, including the demand of coping with time series of variable length and different frame-rates (McLaughlin et al., 2016).

(McLaughlin et al., 2016), presented a model for video-based person re-identification that used color and optical flow pixel information as input to the network to model the temporal structure in the video. To exploit the rich sequence information, (Wang et al., 2014) mainly focus on using the key frame representation. In this approach, since only one fragment is selected to represent the whole sequence, richer information contained in the rest of the sequences is not fully utilized. In (Karanam et al., 2015) and (Zheng et al., 2015) approaches based on feature fusion/encoding have been proposed which exploit the bag-of-words framework to encode a set of frame-wise features into a global vector, but ignore the informative spatio-temporal information of human sequence. Furthermore, some works on video-based person re-identification such as the proposed method in (Wang et al., 2014) and (Liu et al., 2015), can be viewed as extracting low-level 3D features (e.g., HOG3D (Klaser et al., 2008), (Wang et al., 2014), color/gradient features over color channels (Liu et al., 2015)), frame by frame through pre-aligned sequences and aggregating these features afterwards. The approach in (Liu et al., 2015) is taking into account the gait (i.e., the way a person walks) information in a walking cycle specifically for re-identification of walking pedestrians in video frames. The authors in (Wang et al., 2014), proposed a discriminative video fragments selection model by selecting and matching more reliable features from video fragments. However, the temporal sequence nature of videos is not explicitly modeled in these approaches.

2.4 Object Importance Estimation

Finding the important road-users in the scene is crucial in self-driving cars and driver assistance systems in order to interact with other road users and have a safe navigation. Recently many efforts have been devoted to development of vehicles with higher level of autonomy based on scene understanding. For instance, driver's gaze has been wildly studied for determining saliency map and intention prediction relying only on fixation maps (Pugeault and Bowden, 2015). (Underwood et al., 2011) inspects the driver's attention specifically towards pedestrians and motorbikes, and exploits object saliency. In (Palazzi et al., 2017), a computer vision based model is proposed to predict saliency by conducting a data-driven study on drivers' gaze fixations. However, driver's gaze is not always a valid indication of saliency since the driver might look at many unimportant objects in the scene as well.

Different from our proposed method, prediction of important objects is also studied by (Kuen et al., 2016; Li et al., 2016). (Kuen et al., 2016) uses recurrent attention and convolutional-deconvolutional network to tackle the salient object detection problem. Furthermore, the proposed model in (Li et al., 2016) takes a strategy for encoding the underlying saliency prior information, and then sets up a multi-task learning scheme for exploring the intrinsic correlations between salient object detection and semantic image segmentation. However, these methods are not applicable to road user importance estimation in driving scenario which highly depends on the ego car's intention and its interaction with other road users.

Another approach for solving the saliency estimation problem in autonomous driving is using sensor-based methods. LiDAR (Halterman and Bruch, 2010), radars, lasers and sonars (Park et al., 2003) are popular sensors to detect surrounding objects in autonomous systems. For instance, (Sheu et al., 2007) uses smart antennas and proposes a distance awareness system for important object estimation. The model proposed by (Chen et al., 2017) combines the front view of the LiDAR point cloud with region-based features from the bird's eye view for 3D object detection. However, the salient objects are not necessarily the nearest object (e.g. nearest object like a parked car may not pose as much a threat as

a pedestrian intending to cross ego-vehicle's path further down the road). Therefore, visual information is essential for practical autonomous driving systems. For more details about the history of using different sensors and methods for autonomous driving systems please refer to (Janai et al., 2017).

Different from recent works based on estimating a general saliency map of the scene (i.e., a heat map which gives each pixel a relative value of its level of saliency), our proposed method is able to specifically estimate the importance level of all the road users based on the scene context and ego car's intention. Furthermore, unlike the works based on estimating the driver's gaze fixation map, in this work we propose a road user importance estimation method based on human-centric importance annotation.

2.5 Meta-Learning for Few-Shot Learning

A meta-learning model is trained over a variety of learning tasks and optimized for the best performance on a distribution of unseen tasks. This differs from standard machine learning techniques, which involve training on a single task and testing on held-out examples from that task. Please refer to https://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/ for a simple description of meta-learning or learning to learn.

Few-shot classification is an instantiation of meta-learning in the field of supervised learning. Recently there has been a resurgence of interest in few-shot learning based on meta-learning (Finn et al., 2017; Vinyals et al., 2016; Snell et al., 2017; Ravi and Larochelle, 2016; Santoro et al., 2016; Munkhdalai and Yu, 2017; Sung et al., 2018). The existing meta-learning models for few-shot classification can be divided into three types: the learning to fine-tune based, RNN based, and metric learning based. For instance, in (Finn et al., 2017) the MAML model aims to meta-learn an initial condition that is good for fine-tuning on few-shot problems. The model in (Ravi and Larochelle, 2016) is an LSTM-based optimizer that is trained to be specifically effective for fine-tuning. In (Santoro et al., 2016), a recurrent neural network iterates over examples of given problem and accumulates the knowledge required to solve that problem in its hidden activations. However, these recent works either require fine-tuning the target problem (Finn et al., 2017; Ravi and Larochelle, 2016), or need the

use of complex recurrent neural network (RNN) architectures (Santoro et al., 2016; Vinyals et al., 2016), or are based on complicated inference steps (Fei-Fei et al., 2006). In our work, the model is simple and fast and does not need any additional process such as fine tuning. Moreover, we avoid the complexity of recurrent networks, and the issues involved in ensuring the adequacy of their memory. Instead our proposed approach is defined entirely with feed forward convolution neural networks.

The metric based few-shot learning has attracted a lot of interests recently (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018). The basic idea is to learn a metric which can map similar samples close and dissimilar ones distant in the metric space so that a query can be easily classified. Various metric based methods such as siamese networks (Chopra et al., 2005), matching networks (Vinyals et al., 2016), prototypical networks (Snell et al., 2017), and relation networks (Sung et al., 2018) have been proposed. They differ in their ways of learning the metric. For instance, very recently the relation network (Sung et al., 2018) proposed to replace the fixed metric learning part (e.g., Euclidean distance) of the previous works with a deep metric for comparing the relation between images.

The success of metric based methods relies on learning a discriminative metric space. The proposed method in this research can be categorized as the metric learning based framework. To reach the full potential of metric based few-shot learning, we augment the classification loss with a structure-based deep metric learning regularization which enforces the model to map the samples in the support set to well separated clusters in the embedding space. This regularization is based on an improved version of deep metric learning framework in (Oh Song et al., 2017) with no need of sample selection and greedy algorithm. Unlike the metric learning methods based on contrastive (Chopra et al., 2005) or triplet (Schroff et al., 2015) loss that are defined in terms of data pairs or triplets, our approach takes into account the global structure of the embedding space. In fact, the structured margin term in the loss function measures the quality of clustering the data by taking into account the relationship between all the data points in the mini batch at once (instead of data pairs or triplets). Furthermore, this deep learning based metric learning framework does not require the training data to be preprocessed in rigid paired or triplet format and uses a structured prediction framework (Tsochantaridis et al., 2004; Joachims et al., 2009) to ensure that

the score of the ground truth clustering assignment is higher than the score of any other clustering assignment.

Taking advantage of contextual information in the support set is critical in episode-based few-shot learning models. A framework for context modeling in the support set was proposed in (Vinyals et al., 2016) based on a bi-directional LSTM. However, as the number of classes and shots increases, the model is required to learn longer and more complex dependencies, which negatively affects both generalization and efficiency. Furthermore, it imposes an arbitrary ordering on the support set by using bi-directional LSTM (i.e., the embedding changes if we shuffle the support set samples). Moreover, the meta-learner architecture proposed in (Mishra et al., 2017) combines temporal convolutions (which aggregate contextual information from past) with causal attention which pinpoints to specific pieces of information. In this research, we propose a simpler but effective context-aware query embedding framework based on attention mechanism and 1-D CNN for taking into account the context of the support set and its relationship (i.e., task) with query embedding. The proposed query encoder makes the query embedding task-dependent which helps learning a meta-learner with higher generalization power.

Chapter 3

Object Recognition in Distributed Camera Networks

3.1 Introduction

Because the smart camera networks have become increasingly more affordable and perform better in balancing the computational power and energy efficiency, they have been employed in many surveillance tasks including distributed object recognition (Redondi et al., 2015), (Christoudias et al., 2008), (Ferrari et al., 2004), cross view action recognition (Zheng et al., 2016) and person re-identification (Lisanti et al., 2015), (Ahmed et al., 2015) to name just a few.

However, a major challenge in visual sensor networks is limitation in terms of transmission bandwidth, storage and processing power. In the traditional system design for visual sensor networks, images are acquired and compressed locally at the camera nodes, and then transmitted to the base station which performs the specific analysis tasks (e.g., video surveillance, object recognition, etc.). However, recently, a new paradigm has emerged based on analyze-then-compress, where the visual content is processed locally at the camera nodes, to extract a concise representation constituted by local visual features (e.g., SIFT, SURF, HOG). Such features are then compressed and transmitted to the base station for further analysis. Since the feature-based representation is usually more compact than the

pixel-based representation, the analyze-then-compress approach is particularly attractive for those scenarios for which the bandwidth is scarce (Redondi et al., 2015).

Recently several works have been done towards the analyze-then-compress feature compression. For instance, in (Mitra et al., 2014) and (Naikal et al., 2010), authors explore approaches for scalability in large-scale camera networks using recent advances in Compressive Sensing (CS). In (Naikal et al., 2011b), the dimension of the features is reduced using an approach based on Sparse Principal Component Analysis (SPCA). Another study (Turcot and Lowe, 2009), further considered using robust structure-from-motion techniques (e.g., RANSAC) to select strong object features between two camera views, and subsequently rejecting weak features from the final stage of object recognition. However, such solutions are known to break down easily when the camera transformation is large or when the features are extracted from low-quality images. Moreover, most of the existing approaches (e.g., (Turcot and Lowe, 2009), (Christoudias et al., 2008)) require the communication between the smart cameras for selecting the best visual features.

The contributions of this work are as follows:

- First, we propose a method for finding a compact representation of the training set and paying attention to the most important part of the data. We propose a probabilistic algorithm based on the divergence between the probability distributions of the visual features in order to select the most informative visual features and building a compact and physically meaningful model of the training set.

- Second, we introduce an NMF-based scheme for calculating the low-dimensional codes for visual feature of each image, before its transmission to the base station and without any communications between the cameras.

- Third, we elaborate on the distributed recognition task and illustrate the performance of the proposed approach based on the experiments on two challenging and low-resolution multi-view datasets.

The remainder of this chapter is organized as follows. Section 3.1.1 explains the baseline of our proposed distributed object recognition scheme. Section 3.2 elaborates on the proposed

method for obtaining a compact model of the feature histograms in the off-line training stage and then Section 3.2.2 introduces a scheme which uses this compact representation of the training set to encode the histogram of features to low-dimensional codes. Section 3.3 describes the experimental setting, followed by detailed discussion and comparison of the results. The final section concludes the project.

3.1.1 Distributed Object Recognition-The Baseline

SIFT-like feature descriptors have gained major popularity in object recognition task in recent years. These high-dimensional descriptors (e.g., SIFT:128-D, SURF:64-D) are invariant to scale and rotation and therefore are favorable in multi-view object recognition task. In this work, dense SIFT feature descriptors are computed at a grid of overlapped patches in the image. These invariant features are further quantized to form a dictionary of visual words using bag-of-words (BoW) approach (Lee, 2008). Using hierarchical k-means, all the feature descriptors are clustered into visual words. Then a term-frequency inverse-document-frequency (tf-idf) weighted visual histogram is defined for each image (Nister and Stewenius, 2006). Each image histogram is a 1000-D vector and is calculated for all the training and testing images.

In the baseline scenario (without performing any feature selection and compression) the feature histogram of the test object is sent to the base station and a nearest neighbor search is conducted in the training set to find the closest histogram for the testing sample. The class label of the closest histogram in the training set is then used to label the histogram of the test data. In order to fasten the search procedure, the hierarchy clustering using k-means is adopted. We cluster the data into a hierarchical of clusters and do a depth-first search to find the approximate nearest-neighbor to the query point. The chi-square distance is utilized as closeness measure of two l_1-normalized histograms in our Nearest Neighbor classifier. Figure 3.1 shows a simple illustration of this process.

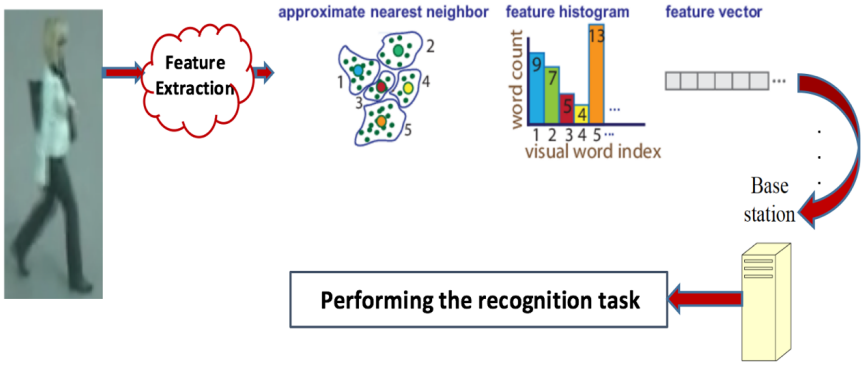


Figure 3.1: Distributed Recognition - Baseline

3.2 Compact Representation of the Features

In this section we elaborate on the proposed method for finding a compact representation of the features by paying attention to the most informative part of the data.

3.2.1 Attention-based Feature Selection

In recent years several studies have been carried out in the context of compact dictionary learning (Jiang et al., 2013; Kong and Wang, 2012; Jiang et al., 2012; Song et al., 2016; Taalimi et al., 2016b), as an approach for finding a compact representation of the data. However, the lack of physical interpretation of the compact dictionary (i.e., physical meaning of each basis in the dictionary) has been a critical shortcoming of the standard dictionary learning techniques. Inspired by the methods based on non-negative matrix factorization, in this part of the project, we address this issue by proposing a novel probabilistic approach for selecting a group of features as a compact representation of all the features in the training set. In other words the goal is to pay "attention" to the most representative part of the data and ignore the rest of it. We believe that nothing is more meaningful for representing the data than the data itself.

Assume there are c classes in the training set and there are N feature histograms $\boldsymbol{h}_i \in R^m$ in each class (i.e., $H = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\} \in R^{m \times N}$). When each bin of histogram is divided by the number of visual words in each cluster, the probability density function (pdf) which represents a probability distribution is produced. Therefore, for each class we have N pdfs as: $F = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N\} \in R^{m \times N}$.

The objective of this stage of the proposed approach is to compare all the *pdf*s in each class and select a few informative ones by solving the following optimization problem:

$$\begin{aligned} &\min_{w_{ij}} \sum_{j=1}^{N}\sum_{i=1}^{N}(\sum_{k=1}^{m}((\boldsymbol{f}_i(k)-\boldsymbol{f}_j(k))\ln(\frac{\boldsymbol{f}_i(k)}{\boldsymbol{f}_j(k)}))w_{ij} \\ &s.t.\ \sum_{i=1}^{N} w_{ij}=1,\ \forall j;\quad (\sum_{i=1}^{N}(\sum_{j=1}^{N}|w_{ij}|^q)^{p/q})^{1/p}\leq\lambda \\ &w_{ij}\geq 0,\ \forall i,j, \end{aligned} \tag{3.1}$$

where $\sum_{k=1}^{m}((\boldsymbol{f}_i(k)-\boldsymbol{f}_j(k))\ln(\frac{\boldsymbol{f}_i(k)}{\boldsymbol{f}_j(k)})$ is the symmetric form of the KL divergence (Kullback and Leibler, 1951). This term measures the difference between all the probability distribution pairs in F.

w_{ij} is defined as the probability of $\boldsymbol{f}_i$ being a representative for $\boldsymbol{f}_j$ (i.e., $w_{ij}\in[0,1]$). Therefore, we must have $\sum_{i=1}^{N} w_{ij}=1$, to assure that the probability of each $\boldsymbol{f}_j$ being represented via $F=\{\boldsymbol{f}_1,\ldots,\boldsymbol{f}_N\}$ is equal to one. Hence, the first term in Eq. 3.1 is the cost of representing $\boldsymbol{f}_j$ via $\boldsymbol{f}_i$, which is defined as the divergence measure between them, times the probability of the occurrence of this event. We define $\boldsymbol{W}\in R^{N\times N}$ as the probability matrix for all the $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ pairs (i.e., w_{ij} is the ith row and jth column entry of the $\boldsymbol{W}$ matrix). In other word, when $\boldsymbol{f}_i$ is a representative for $\boldsymbol{f}_j$, the corresponding row in the $\boldsymbol{W}$ matrix is non-zero. Since our goal is to find some few representations of $\boldsymbol{f}_j$ using $\boldsymbol{f}_i$, we impose a row-sparsity constraint on the $\boldsymbol{W}$ matrix in order to select only a few $\boldsymbol{f}_i$s and set the other rows of $\boldsymbol{W}$ entirely equal to zero.

In order to achieve this goal, we exploit a joint $l_{p,q}$ norm regularization (the second constraint in Eq. 3.1), where λ is a regularization parameter and determines the number of non-zero rows of the $\boldsymbol{W}$ matrix. The l_{pq} norm, is convex for $p\geq 1$ and $q\geq 1$; otherwise it is a quasi-norm and is non-convex. In fact, we can consider any $q\geq 1$, however, $l_{p,\infty}$ (i.e., $q=\infty, p\leq 1$) has the property of giving us the real number of non-zero features which is the desired goal in our feature selection task. The $l_{p,\infty}$ penalty is a convex relaxation of a pseudo-norm which counts the number of non-zero rows in $\boldsymbol{W}$. Another consideration in l_{pq} norm is the choice of p. Some works such as Chartrand and Staneva (2008) have investigated the l_p norm with $0<p<1$. It is worth noting that for $0<p<1$, Eq. 3.1 is not a convex

problem and we cannot guarantee the global minimum and the solution is not unique and it highly depends on initialization. In other words, even though $0 < p < 1$ might lead to more sparse result, but the solution would not be consistent. Additionally, choosing an optimum initialization method is not straight forward. Hence, in this work, we consider $p = 1$ that leads to a convex problem and global minimum for the optimization problem in Eq. 3.1. As a result, the proposed optimization problem will have the following form:

$$\begin{aligned} &\min_{w_{ij}} \sum_{j=1}^{N}\sum_{i=1}^{N}(\sum_{k=1}^{m}((\boldsymbol{f}_i(k)-\boldsymbol{f}_j(k))\ln(\frac{\boldsymbol{f}_i(k)}{\boldsymbol{f}_j(k)}))w_{ij} \\ &s.t.\ \sum_{i=1}^{N} w_{ij}=1,\ \forall j; \quad \sum_{i=1}^{N}(\max_{1\le j\le N} w_{ij}) \le \lambda \\ &w_{ij}\ge 0,\ \forall i,j, \end{aligned} \tag{3.2}$$

We refer to Eq. 3.2 as the Divergence-based Feature Selection (DFS) method. We select the feature histograms, corresponding to indices of non-zero rows of $\boldsymbol{W}$ as our representative features in each class and we repeat this process for all the classes in the training set. The number of selected features for each class is determined by the regularization parameter λ (i.e., λ is roughly the number of non-zero rows in the $\boldsymbol{W}$ matrix). It is important to note that the value of λ should satisfy $\lambda \le N$ (i.e., N is the number of training data in each class), otherwise the $\boldsymbol{W}$ matrix would be the identity matrix, since each probability distribution $\boldsymbol{f}_i$ is the best representation for itself. The convex optimization problem in Eq. 3.2 is solved using the Alternating Direction Method of Multipliers framework in Boyd et al. (2011).

3.2.2 Generating Low Dimensional Feature Codes

After constructing a compact representation of the feature histograms for all the classes in the training set, it will be saved in the smart cameras' memory (we refer to this compact representation as $\boldsymbol{D}$). In the on-line testing stage in each camera, a feature histogram $\boldsymbol{h}_i$ is extracted for the ith test image at each of the p local cameras independently, and we calculate the corresponding code (i.e., $\boldsymbol{s}_i$) for each feature histogram, using a supervised constrained non-negative matrix factorization scheme:

$$\min_{\boldsymbol{s}_i}\{\|\boldsymbol{h}_i - \boldsymbol{D}\boldsymbol{s}_i\|_2^2\}, \quad i = (1, ..., p),$$
$$s.t. \;\; \boldsymbol{s}_i(j) \geq 0, \; j = 1, \ldots, k, \qquad \sum_{j=1}^{k} \boldsymbol{s}_i(j) = 1 \tag{3.3}$$

where $\boldsymbol{h}_i \in R^{m\times 1}$, $\boldsymbol{D} \in R^{m\times k}$, $\boldsymbol{s}_i \in R^{k\times 1}$ and $k << m$. k is the number of features that have been selected for the whole training set in the previous step (i.e., number of columns of $\boldsymbol{D}$), and m is the dimension of the original feature histograms (i.e., 1000 in our setting). The optimization problem in Eq. 3.3 is developed from the Non-negative Constrained Least Squares (NCLS) method (Chang and Heinz, 2000) in conjunction with the sum-to-one constraint. The objective is to minimize the least squares error:

$$\min_{\boldsymbol{s}_i} \left\|\hat{\boldsymbol{h}}_i - \hat{\boldsymbol{D}}\boldsymbol{s}_i\right\|_2^2, \quad i = (1, ..., p),$$
$$s.t. \;\; \boldsymbol{s}_i(j) \geq 0, \; j = 1, \ldots, k, \tag{3.4}$$

where $\hat{\boldsymbol{h}}_i$ and $\hat{\boldsymbol{D}}$ are the augmented matrices

$$\hat{\boldsymbol{h}}_i = \begin{bmatrix} \delta\boldsymbol{h}_i \\ \mathbf{1} \end{bmatrix}, \quad \hat{\boldsymbol{D}} = \begin{bmatrix} \delta\boldsymbol{D} \\ \mathbf{1}^T \end{bmatrix} \tag{3.5}$$

with δ being a small weight and $\mathbf{1}^T$ is a row vector of all 1s. This augmentation is used to incorporate the sum-to-one constraint. The constrained minimization problem in Eq. 3.4 is solved by a standard active set method (Bro and De Jong, 1997). This process is simple and can be done fast inside each smart camera. After finding $\boldsymbol{s}_i \in R^{k\times 1}$ in each camera ($i = 1, \ldots, p$), these low dimensional codes will be sent to the base station for performing the intended recognition task. Transmitting codes with k dimension instead of feature histograms with m dimension leads to major saving in bandwidth of the wireless network (the compression ratio: m/k, $k << m$) as well as better recognition accuracy.

3.3 Experiments and Results

In this work, we validate our proposed feature encoding scheme on two multi-view recognition tasks including pedestrian recognition in surveillance video and distributed object recognition in smart camera networks.

3.3.1 Datasets

Person Re-ID (PRID) dataset

The Person Re-ID (PRID) dataset (Hirzer et al., 2011) is one of the few multi-view datasets which includes multi image frames for each pedestrian recorded from two different, static surveillance cameras. Images from these cameras contain a viewpoint change and a stark difference in illumination, background and camera characteristics. Since images are extracted from trajectories, several different poses per pedestrian are available in each camera view. It contains recorded frames of 475 person trajectories from one view and 856 from the other one, with 245 persons appearing in both views (Hirzer et al., 2011). Figure 3.2 illustrates some sample frames of this dataset.

Berkeley Multi-view Wireless

We also employ Berkeley Multi-view Wireless (*BMW*) database (Naikal et al., 2010) in order to evaluate the performance of our proposed algorithm on real multi-view object recognition. It is important to note that the image quality in this database is considerably lower than many existing high-resolution databases, which is intended to reproduce realistic imaging conditions for surveillance applications (Naikal et al., 2010). This fact makes the recognition more difficult. *BMW* consists of multiple-view images of 20 landmark buildings on the Berkeley campus. Few samples of images in this database are shown in Figure 3.3. For each building, wide-baseline images were captured from 16 different vantage points. Further, at

Figure 3.2: PRID dataset samples

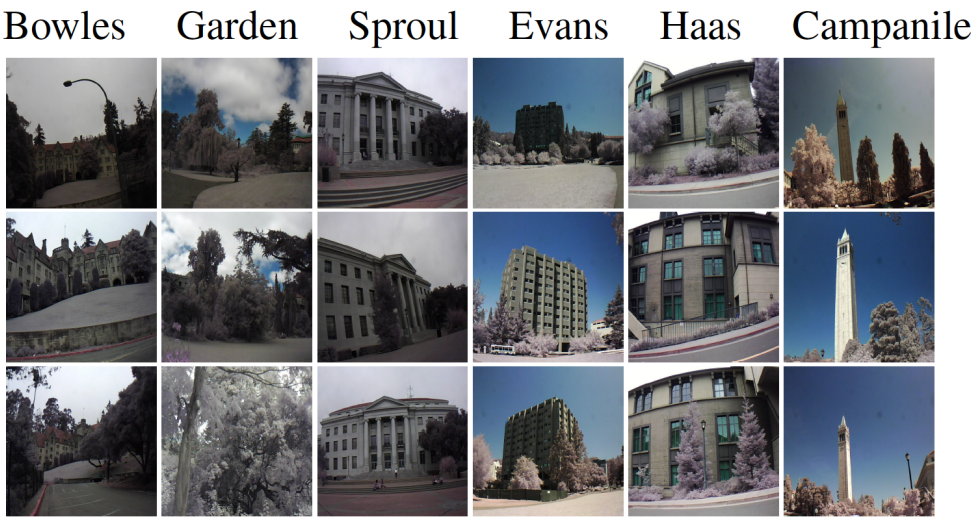


Figure 3.3: Few sample images of BMW database

each vantage point, 5 short-baseline images were taken by five camera sensors simultaneously, thereby summing to 80 images per category (Naikal et al., 2010). All images are 640×480 RGB color images. We divide the database into a training set and a testing set. Between 5 cameras, images from camera #2 captured at the even vantage points of each category are assigned as the training set, and the remaining images are assigned to the testing set (Naikal et al., 2011a).

3.3.2 Pedestrian Recognition in Surveillance Video

In the first experiment on PRID dataset, we consider 30 different frames for each pedestrian in each camera and we randomly choose 20 persons in the dataset for the recognition task. Hence, there are 1200 images for which we randomly pick half of it as the training set and the rest as the testing set. The dimension of the original feature histograms is 1000. Figure 3.4 shows the recognition accuracy versus the compression rate using the proposed DFS method.

In this figure, we can observe that with compression rate of 2.94 (i.e., features with dimension of 340), the accuracy is slightly better than using the original features. The reason is that our feature selection scheme omits those features which are closer to the features from other classes than the features in their own class. We define the compression ratio as dimension of the original features divided by dimension of the encoded features (i.e.,

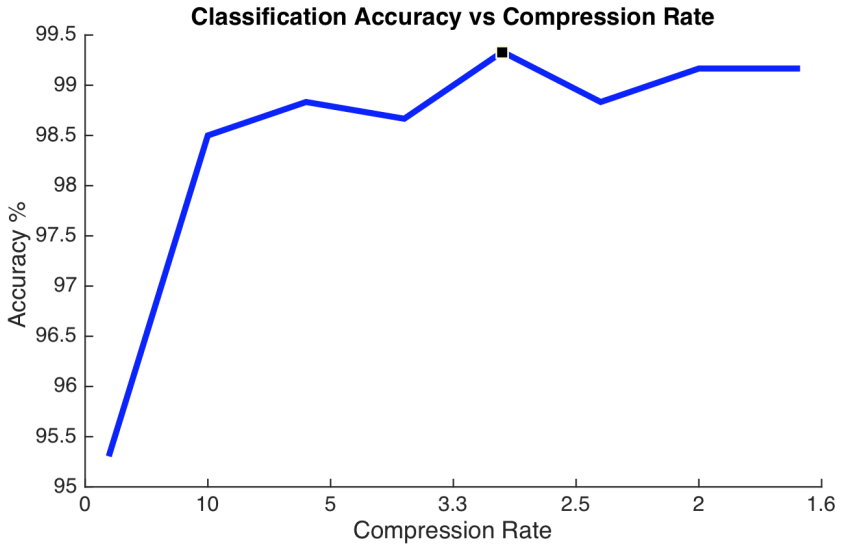


Figure 3.4: Recognition accuracy for different dimensions of the feature histograms using the DFS method

k in Eq. 3.3). For instance, in Figure 3.4 at the marked point on the curve with compression ratio of 2.94, the feature dimension is equal to $1000/2.94 = 340$. Figure 3.5 illustrates the recognition accuracy for classification of all the 20 persons in the experiment with 340-D feature histograms ($\lambda = 17$ in Eq. 3.2). It is worth noting that the recognition task in this experiment is different from person re-identification task which is based on image matching and retrieval.

3.3.3 Building Recognition in BMW Dataset

In the second experiment (on BMW), there exist 16 different vantage points, and at each vantage point, images are taken by five cameras simultaneously, thereby summing to 80 images per category. In this section, we compare the recognition accuracy of DFS method with two other existing works. Table 3.1 demonstrates the classification accuracy of different methods based on Sparse PCA (SPCA) (Naikal et al., 2011b) and Structure from Motion (SfM) (Turcot and Lowe, 2009). To have a fair comparison, we set up the same experimental environment as the other two works. In fact, we only considered 8 images (even vantage points of camera #2) from each object for training and the rest of images from other cameras for testing (and compression ratio: 2.4).

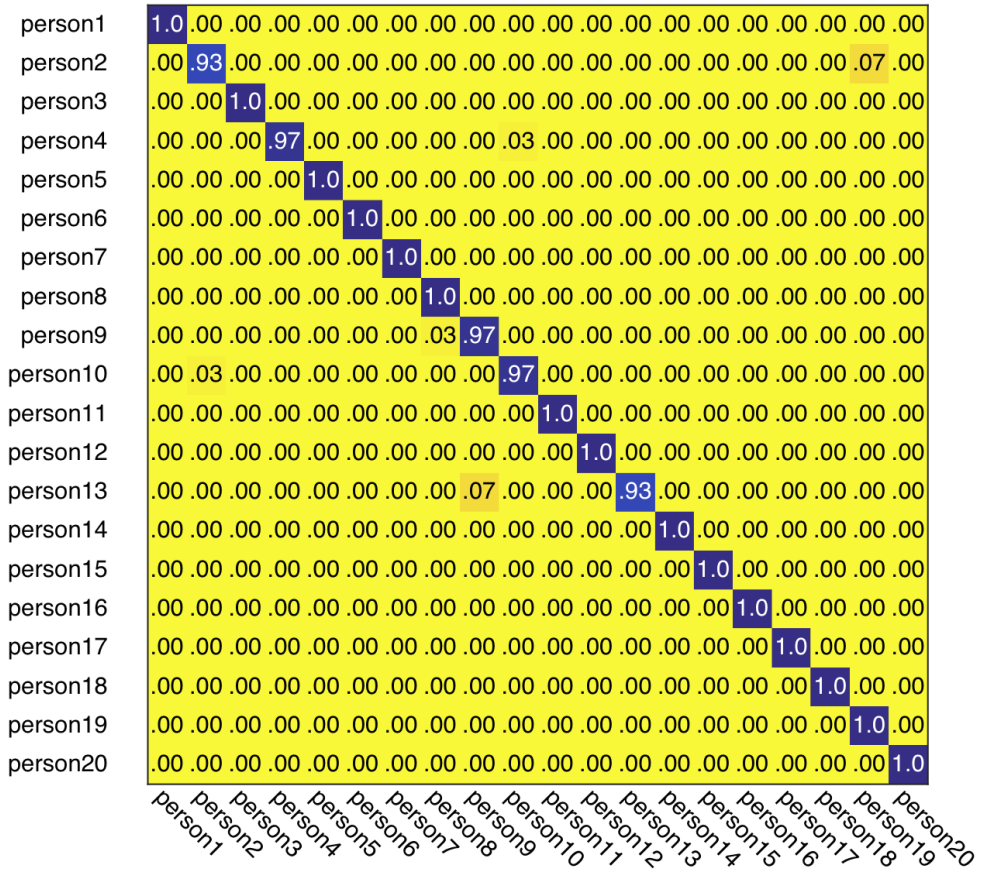


Figure 3.5: Confusion matrix of pedestrian recognition via DFS on PRID dataset using 340-D features.

Table 3.1: Comparison of recognition accuracy rate of different methods. For all methods compression ratio is set to 2.4 : 1.

Class	Baseline(%)	DFS(%)	SPCA(%)	SfM (%)
1	97.01	**99.00**	94.44	83.33
2	85.00	86.34	**91.66**	90.27
3	79.04	**86.00**	66.66	58.33
4	97.68	**100.00**	81.94	65.27
5	85.05	86.03	**91.66**	81.94
6	88.03	**93.23**	88.88	87.50
7	92.00	**98.00**	93.05	86.11
8	**99.00**	94.01	91.66	72.22
9	89.79	**91.56**	73.61	63.88
10	74.38	**76.43**	65.27	61.11
11	82.33	**96.52**	76.38	69.44
12	85.94	**99.25**	83.33	70.83
13	**92.50**	86.00	72.22	52.77
14	92.61	**100.00**	93.05	90.27
15	83.44	**86.01**	80.55	75.00
16	83.00	**100.00**	79.16	80.55
17	**100.00**	**100.00**	90.27	84.72
18	94.61	**100.00**	93.05	**100.00**
19	89.22	**100.00**	83.33	86.11
20	90.61	93.05	**100**	95.83
Avg.	88.68	**93.50**	84.51	77.77

For most of the object categories our proposed method, outperforms SPCA and SfM based approaches. One important reason for outperforming the proposed DFS method compared to other two methods is that in contrast to SPCA and SfM methods, the physical interpretation of the reduced space is preserved during the dimensionality reduction procedure which is critical in the recognition task.

3.4 Summary

In this work, we proposed a probabilistic encoding approach based on divergence of the probability distributions of the visual features in limited bandwidth distributed camera networks. The performance of the proposed approach was discussed in two surveillance recognition tasks. The proposed DFS approach is applicable to the variety of distributed computer vision tasks based on transmission of the visual features in a network (e.g., cross view action recognition, person re-identification, etc.).

Chapter 4

Attention-based Person Re-identification in Distributed Camera Networks

4.1 Introduction

Despite recent attempts for solving the person re-identification problem, it remains a challenging task since a person's appearance can vary significantly when large variations in view angle, human pose and illumination are involved. The concept of attention is one of the most interesting recent architectural innovations in neural networks. Inspired by that, in this research we propose a novel approach based on using a gradient-based attention mechanism in deep convolution neural network for solving the person re-identification problem. Our model learns to focus selectively on parts of the input image for which the networks' output is most sensitive to.

Recently, person re-identification has gained increasing research interest in the computer vision community due to its importance in multi-camera surveillance systems. Person re-identification is the task of matching people across non-overlapping camera views. A typical re-identification system takes as input two images of person's full body, and outputs either a similarity score between the two images or the decision of whether the two images belong

to the same identity or not. Person re-identification is a challenging task. In fact, different individuals can share similar appearances and also appearance of the same person can be drastically different in two different views due to several factors such as background clutter, illumination variation and pose changes.

It has been proven that humans do not focus their attention on an entire scene at once when they want to identify another person (Xu et al., 2015). Instead, they *pay attention* to different parts of the scene (e.g., the person's face) to extract the most discriminative information. Inspired by this observation, we study the impact of attention mechanism in solving person re-identification problem. The attention mechanism can significantly reduce the complexity of the person re-identification task, where the network learns to focus on the most informative regions of the scene and ignores the irrelevant parts such as background clutter. Exploiting the attention mechanism in person re-identification task is also beneficial at scaling up the system to large high quality input images.

With the recent surge of interest in deep neural networks, attention based models have been shown to achieve promising results on several challenging tasks, including caption generation (Xu et al., 2015), machine translation (Bahdanau et al., 2014) and object recognition (Ba et al., 2014). However, attention models proposed so far, require defining an explicit predictive model, whose training can pose challenges due to the non-differentiable cost. Furthermore, these models employ Recurrent Neural Network (RNN) for the attention network and are computationally expensive or need some specific policy algorithms such as REINFORCE (Ba et al., 2014; Williams, 1992) for training.

In this research, we introduce a novel model architecture for person re-identification task which improves the matching accuracy by taking advantage of attention mechanism. The contributions of this research are the following:

- We propose a CNN-based task-driven attention model which is specifically tailored for the person re-identification task in a triplet architecture. Our model generates highly discriminative features by fusion of global and local features which are trained based on two losses.

- The network is computationally efficient during inference, since it first finds the most discriminative regions in the input image and then performs the deep CNN feature extraction only on these selected regions.

- Finally, we qualitatively and quantitatively validate the performance of our proposed model by comparing it to the state-of-the-art performance on three challenging benchmark datasets: CUHK01(Li and Wang, 2013), CUHK03 (Li et al., 2014) and Market 1501 (Zheng et al., 2015).

Generally, existing approaches for person re-identification are mainly focused on two aspects: learning a distance metric (Liao et al., 2015; Pedagadi et al., 2013; Su et al., 2015) and developing a new feature representation (Varior et al., 2016c; Zhao et al., 2013b; Liao et al., 2010; Ojala et al., 2002; Zhao et al., 2013a; Zheng et al., 2015). In distance metric learning methods, the goal is to learn a metric that emphasizes inter-personal distance and de-emphasizes intra-person distance. The learnt metric is used to make the final decision as to whether a person has been re-identified or not (e.g., KISSME (Köstinger et al., 2012), XQDA (Liao et al., 2015), MLAPG (Su et al., 2015) and LFDA (Pedagadi et al., 2013)). In the second group of methods based on developing new feature representation for person re-identification, novel feature representations were proposed to address the challenges such as variations in illumination, pose and view-point (Varior et al., 2016c). The Scale Invariant Local Ternary Patterns (SILTP) (Liao et al., 2010), Local Binary Patterns (LBP) (Ojala et al., 2002), Color Histograms (Zhao et al., 2013a) or Color Names (Zheng et al., 2015) (and combination of them), are the basis of the majority of these feature representations developed for human re-identification.

In the recent years, several approaches based on Convolutional Neural Network (CNN) architecture for human re-identification have been proposed and achieved great results (Cheng et al., 2016a; Li et al., 2014; Ahmed et al., 2015). In most of the CNN-based approaches for re-identification, the goal is to jointly learn the best feature representation and a distance metric (mostly in a Siamese fashion (Bromley et al., 1993)). With the recent development of RNN networks, the attention-based models have demonstrated outstanding performance on several challenging tasks including action recognition (Sharma et al., 2015).

At the time of writing this research, except for one recent work (Liu et al., 2016), the attention mechanism has not yet been studied in the person re-identification literatures. In (Liu et al., 2016), the RNN-based attention mechanism is based on the attention model introduced in (Sharma et al., 2015) for action recognition.

Different from (Liu et al., 2016), in our model the selection of the salient regions is made using a novel gradient-based attention mechanism, that efficiently identifies the input regions for which the network's output is most sensitive to. Moreover, our model does not use the RNN architecture as in (Liu et al., 2016), thus is computationally more efficient and easier to train. Furthermore, in (Liu et al., 2016) the attention model requires a set of multiple glimpses to estimate the attention which is not required in our proposed architecture.

4.2 Model Architecture

In this section we introduce our gradient-based attention model within a triplet comparative platform specifically designed for person re-identification. We first describe the overall structure of our person re-identification design, then we elaborate on the network architecture of the proposed attention mechanism.

4.2.1 Triplet Loss

We denote the triplets of images by $< I_i^+, I_i^-, I_i >$, where I_i^+ and I_i are images from the same person and I_i^- is the image from a different person. As illustrated in Figure 4.1, each image initially goes through the global attention network and salient regions of the image are selected (i.e., X^a). Then only these selected regions of the image pass through the local deep CNN. The local CNN network then maps this raw image regions to the feature space $< f_l(X_i^{a+}), f_l(X_i^{a-}), f_l(X_i^a) >$, such that the distance of the learned features of the same person is less than the distance between the images from different persons by a defined margin. Hence, the goal of the network is to minimize the following cost function for N triplet images:

$$J = \frac{1}{N}\sum_{i=1}^{N} max(\left\|f_l(X_i^a) - f_l(X_i^{a+})\right\|_2^2 - \left\|f_l(X_i^a) - f_l(X_i^{a-})\right\|_2^2 + \alpha, 0), \tag{4.1}$$

where α is a predefined margin which helps the model to learn more discriminative features. Choosing the right triplets is critical in training of the triplet loss. For instance, if we use easy negative and positive samples for each anchor, the loss would be zero all the time and the model will not learn anything during training. We define the hard triplets as the triplets where the distance of the negative sample embedding to the anchor embedding is less than the distance of the positive sample embedding to the anchor embedding. We also define semi-hard triplets as triplets that satisfy the following inequality:

$$\left\| f_l(X_i^a) - f_l(X_i^{a+}) \right\|_2^2 < \left\| f_l(X_i^a) - f_l(X_i^{a-}) \right\|_2^2 < \left\| f_l(X_i^a) - f_l(X_i^{a+}) \right\|_2^2 + \alpha \tag{4.2}$$

For training of our model we follow the hard and semi-hard negative sample mining based on the framework proposed in Schroff et al. (2015). It is important to note that the above triplet architecture is used only in the training phase and during testing, the distances between embedding of the query and gallery images are computed and used for ranking.

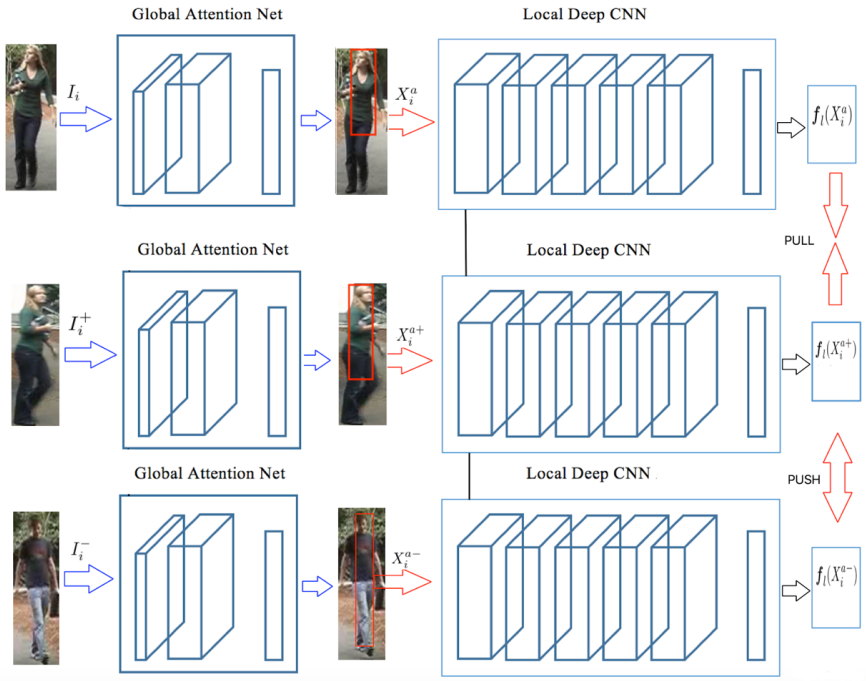


Figure 4.1: The architecture of the proposed Gradient-based Attention Network (GAN) in training phase.

4.2.2 Gradient-based Attention Network

The proposed Gradient-based Attention Network (GAN) is capable of extracting information from an image by adaptively selecting the most informative image regions and only processing the selected regions at high resolution. The whole model comprises of two blocks: the global attention network G and the local deep CNN network L. The global network consists of only two layers of convolution and is computationally efficient, whereas the local network is deeper (e.g., many convolutional layers) and is computationally more expensive, but has better performance.

We refer to the feature representation of the global layer and the local layer by f_g and f_l, respectively. The attention model uses backpropagation to identify the few vectors in the global feature representation $f_g(I)$ to which the distribution over the output of the network (i.e., $\boldsymbol{h}_g$) is most sensitive. In other words, given the input image I, $f_g(I) = \{\boldsymbol{g}_{i,j} | (i,j) \in [1, d_1] \times [1, d_2]\}$, where d_1 and d_2 are spatial dimensions that depend on the image size and $\boldsymbol{g}_{i,j} = f_g(x_{i,j}) \in \mathbb{R}^D$ is a feature vector associated with the input region (i,j) in I, i.e., corresponds to a specific receptive field or a patch in the input image. On top of the convolution layers in attention model, there exists a fully connected layer followed by a max pooling and a softmax layer, which consider the bottom layers' representations $f_g(I)$ as input and output a distribution over labels, i.e., $\boldsymbol{h}_g$.

Next, the goal is to calculate the attention map. We use the entropy of the output vector $\boldsymbol{h}_g$ as a measure of saliency in the following form:

$$H = \sum_{l=1}^{C} \boldsymbol{h}_g^l \log(\boldsymbol{h}_g^l), \tag{4.3}$$

where C is the number of class labels in the training set. In order to find the attention map we then compute the norm of the gradient of the entropy H with respect to the feature vector $\boldsymbol{g}_{i,j}$ associated with the input region (i,j) in the input image:

$$A_{i,j} = \left\| \nabla_{\boldsymbol{g}_{i,j}} H \right\|_2, \tag{4.4}$$

hence, the whole attention map would be $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$ for the whole image. Using the attention map $\boldsymbol{A}$, we select a set of k input region positions (i, j) corresponding to the $A_{i,j}$s with the k largest values. The selected regions of the input image corresponding to the selected positions are denoted by $X^a = \{x_{i,j} | (i, j) \in [1, d_1] \times [1, d_2]\}$, where each $x_{i,j}$ is a patch in input image I. Exploiting the gradient of the entropy as the saliency measure for our attention network encourages selecting the input regions which have the maximum effect on the uncertainty of the model predictions. Note that all the elements of the attention map $\boldsymbol{A}$ can be calculated efficiently using a single pass of backpropagation. For training of the global attention network (G), we maximize the log-likelihood of the correct labels (using cross-entropy objective function).

After selecting the salient patches (X^a) within the input image, the local deep network (L) will be applied only on those patches. This leads to major saving in computational cost of the network and accuracy improvement by focusing on the informative regions of the person's image. The local deep CNN network (L) is trained on attended parts of the input image using the triplet loss introduced in Eq. 4.1. We denote the feature representation created by the local deep network L as $f_l(X^a)$.

In the test time, the local feature representation $f_l(X^a)$ and the global feature representation $f_g(I)$ are fused to create a refined representation of the whole image. In feature fusion, we replace the global features (low resolution features) corresponding to the attended regions (i.e., X^a) with the rich features from the deep CNN (high resolution features). Fusion of the features which are trained based on two discriminative losses leads to highly accurate retrieval performance.

4.3 Experiments and Results

4.3.1 Network Design

We implement our network using TensorFlow (et al., 2015) deep learning framework. The training of the GAN converges in roughly 6 hours on Intel Xeon CPU and NVIDIA TITAN X GPU. In the global attention network (see Figure 6.1), there are 2 convolutional layers, with

7×7 and 3×3 filter sizes, 12 and 24 filters, respectively. On the top of the two convolution layers in the global attention network there are one fully connected layer, a max pooling and a softmax layer. The global attention network is trained once for the whole network with cross-entropy loss. The set of selected patches X^a is composed of eight patches of size 14×14 pixels (experiments showed that the marginal improvement becomes insignificant beyond 8 patches). The Inception-V3 (Szegedy et al., 2016) model pretrained on Imagenet is used for the local deep CNN. Inception-V3 is a 48-layer deep convolutional architecture and since it employs global average pooling instead of fully-connected layer, it can operate on arbitrary input image sizes. The output of the last Inception block is aggregated via global average pooling to produce the feature embedding. We use Batch Normalization (Ioffe and Szegedy, 2015) and Adam (Kingma and Ba, 2014) for training our model. We have employed the same scheme for data augmentation as in (Cheng et al., 2016a). Furthermore, we have used $\alpha = 0.02$ in Eq. 4.1 and exponential learning rate decay for the training (initial learning rate: 0.01).

4.3.2 Datasets

There are several benchmark datasets for evaluation of different person re-identification algorithms. In this research we use CUHK01 (Li and Wang, 2013), CUHK03 (Li et al., 2014) and Market 1501 (Zheng et al., 2015) which are three of the largest benchmark datasets suitable for training the deep convolutional network. The following figures show some sample images from each dataset.

CUHK01 dataset contains 971 persons captured from two camera views in a campus environment. Camera view A captures frontal or back views of a person while camera B captures the person's profile views. Each person has four images with two from each camera. We use 100 persons for testing (Figure 4.2).

CUHK03 dataset contains $13,164$ images of $1,360$ identities. All pedestrians are captured by six cameras, and each person's image is only taken from two camera views. It consists of manually cropped person images as well as images that are automatically detected for simulating more realistic experiment situation. In our experiments we used the cropped person images. We use 100 persons for testing (Figure 4.3).

Figure 4.2: CUHK01 image samples

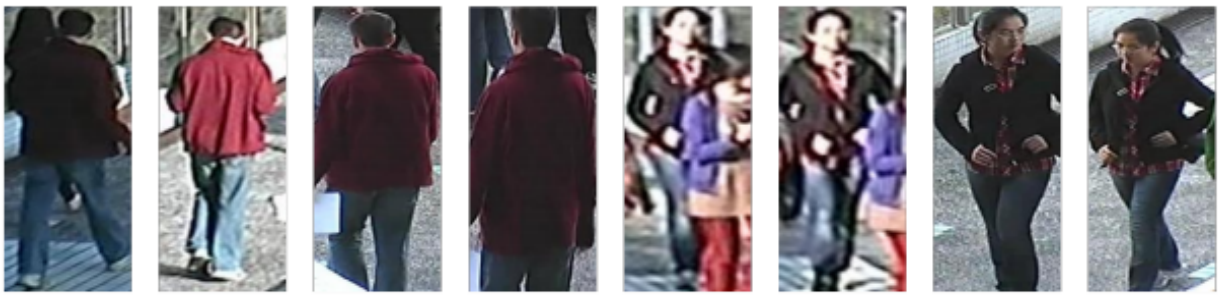


Figure 4.3: CUHK03 image samples

Market1501 dataset contains 32,688 bounding boxes of 1,501 identities, most of which are cropped by an automatic pedestrian detector. Each person is captured by 2 to 6 cameras and has 3.6 images on average at each viewpoint. In our experiments, 750 identities are used for training and the remaining 751 for testing (Figure 4.4).

4.3.3 Evaluation Metric and Results

We adopt the widely used Rank1 accuracy for quantitative evaluations. Also, since the mean Average Precision (mAP) has been used for evaluation on Market 1501 data set in previous works, we use mAP for performance comparison on Market 1501 as well. For datasets with two cameras, we randomly select one image of a person from camera A as a query image and one image of the same person from camera B as a gallery image. For each image in the

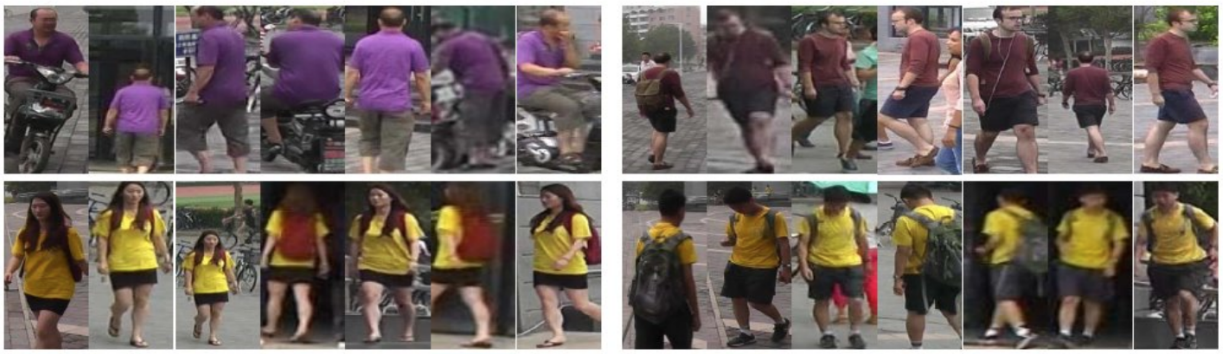


Figure 4.4: Market 1501 image samples

query set, we first compute the distance between the query image and all the gallery images using the Euclidean distance and then return the top n nearest images in the gallery set. If the returned list contains an image featuring the same person as that in the query image at k-th position, then this query is considered as success of rank k. Table 4.1 shows the rank1 accuracy of our model compared to state-of-the-art. It can be observed that the GAN (ours) outperforms all the other methods. Very recently other method (SPReID (Kalayeh et al., 2018)) has been proposed that set a new state-of-the-art. Even though, (Kalayeh et al., 2018) is published after the proposed method in this research, but we include it in our comparisons for completeness (last row in Table 4.1). One success (top) and fail (bottom) case in rank1 retrieval on Market 1501 data set using GAN is shown in Figure 4.5.

Furthermore, GAN is computationally more efficient compared to the case where the local CNN is applied on the whole input image. In practice we observed a time speed-up by a factor of about 2.5 by using GAN (fusion of local and global features) in test stage (tested on 100 test images).

4.3.4 Interpretable Deep Retrieval Model

The visualization of the attention map in our proposed Global attention net is shown in Figure 4.6 and 4.7. These samples are part of the test query samples in Market 1501 dataset that are correctly re-identified by our model. These results show how the network is making its decisions and it thus makes our deep learning model more interpretable.

Table 4.1: Rank1 accuracy (%) comparison of the proposed method to the state-of-the-art.

Method	Market 1501	CUHK01	CUHK03	mAP (Market)
KISSME (Köstinger et al., 2012)	-	29.40	14.12	19.02
GS-CNN (Varior et al., 2016a)	65.88	-	61.08	39.55
DGD (Xiao et al., 2016)	59.53	-	-	31.94
LS-CNN (Varior et al., 2016b)	61.60	-	57.30	35.30
SCSP (Chen et al., 2016)	51.9	-	-	26.35
DNS (Zhang et al., 2016)	55.40	-	-	35.68
Spindle (Zhao et al., 2017a)	76.90	-	88.50	-
P2S (Zhou et al., 2017)	70.72	77.34	-	44.27
PrtAlign (Zhao et al., 2017b)	81.00	88.50	81.60	63.40
PDC (Su et al., 2017)	84.14	-	88.70	63.41
SSM (Bai et al., 2017)	82.21	-	-	68.80
JLML (Li et al., 2017)	85.10	87.00	83.20	65.50
TriNet (Hermans et al., 2017)	84.92	-	-	69.14
Ours	**86.67**	**89.90**	**88.80**	**75.32**
SPReID (Kalayeh et al., 2018)	93.68	-	94.28	84.92

Figure 4.5: Rank1 retrieval on Market 1501. Top figure shows an example of successful retrieval using our model and the bottom figure shows a fail case for rank1. However in the fail case, GAN can still retrieve the image in rank4. left images are the query and the right images are the ranked images using GAN.

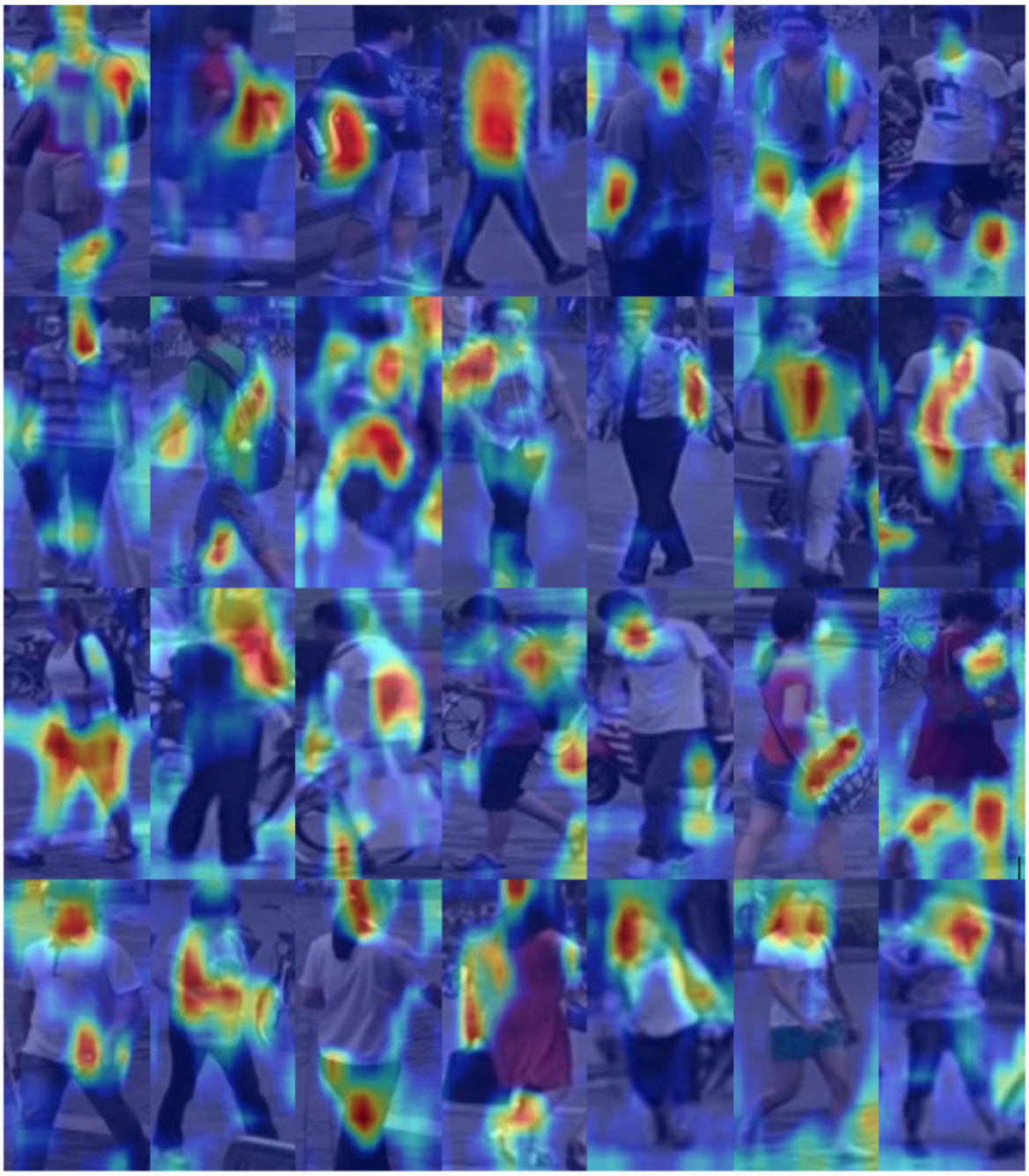


Figure 4.6: Visualization of the attention map produced by our proposed method

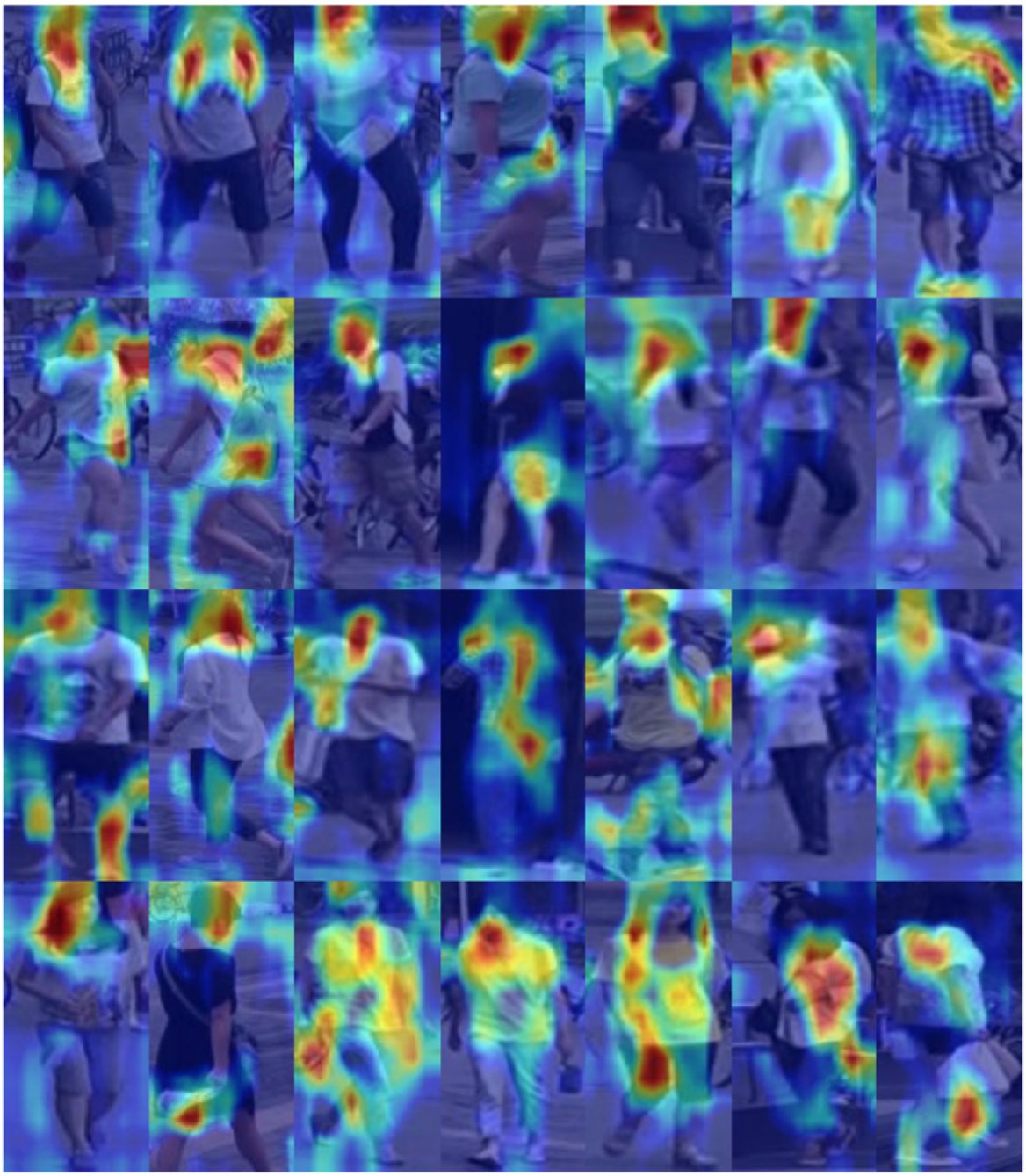


Figure 4.7: Visualization of the attention map produced by our proposed method (2)

For example, the visualization of the results shows how the attention model is able to focus on very detailed and discriminative parts of the input image (e.g., person's face, backpack, shoes, legs, t-shirts, things in their hands) Also, we can observe that by using our attention model, our re-identification system can successfully ignore the background clutter.

4.4 Conclusion

In this research, we introduced an attention mechanism for person re-identification task and we showed how paying attention to important parts of the person's image while still considering the whole image information, leads to highly discriminative feature embedding space and an accurate person re-identification system. Furthermore, thanks to the computational efficiency resulting from the attention architecture, we would be able to use deeper neural networks and high resolution images in order to obtain higher accuracy.

Chapter 5

Context Aware Road-user importance Estimation (iCARE)

5.1 Introduction

Road-users are a critical part of decision-making for both self-driving cars and driver assistance systems. Some road-users, however, are more important for decision-making than others because of their respective intentions, ego-vehicle's intention and their effects on each other. In this research, we propose a novel architecture for road-user importance estimation which takes advantage of the local and global context of the scene. For local context, the model exploits the appearance of the road users (which captures orientation, intention, etc.) and their location relative to ego-vehicle. The global context in our model is defined based on the feature map of the convolutional layer of the module which predicts the future path of the ego-vehicle and contains rich global information of the scene (e.g., infrastructure, road lanes, etc.), as well as the ego-vehicle's intention information. Systematic evaluations of our proposed method against several baselines show promising results.

In real-world driving, at any given time, there can be many road-users in the ego-vehicle's vicinity. Some road-users directly affect ego-vehicle's behavior (i.e. brake, steer), while some could be a potential risk and others who do not pose a risk at this time or in the near future (as illustrated in Figure 5.1). The ability to discern how important or relevant any given road-user is to an ego-vehicle's decision is vital for building trust with human drivers or

passengers, transparency with law makers, promoting human-centric thought process, etc., for both driver assistance systems and self-driving cars. In this research, we propose to estimate road-user importance based on visually guided information.

Given a single image of a driving scene, visually, humans have an unparalleled ability to determine which road-users are affecting or likely to affect the ego-vehicle's behavior. Humans leverage information such as traffic rules, intended path of ego-vehicle, potential trajectory of road participants, location, etc. As many of these information can be inferred from the image, we propose a method to estimate road-user importance by fusion of local and global context. We call this method, Context Aware Road-user Importance Estimation (iCARE).

In iCARE, local context is represented by appearance of road-users (which captures orientation, intention, etc.) and their location relative to ego-vehicle. Global context is represented with the feature map of the last convolutional layer of the model which is trained to predict the future path of the ego car. To some extent this can be considered intention-based context because this same representation can be used to predict future path of ego-vehicle. To this end, the main contributions of this research are as following:

- Designing a new image-based framework for estimating the importance level of road-users which is critical for autonomous driving and advanced driver-assistance systems.

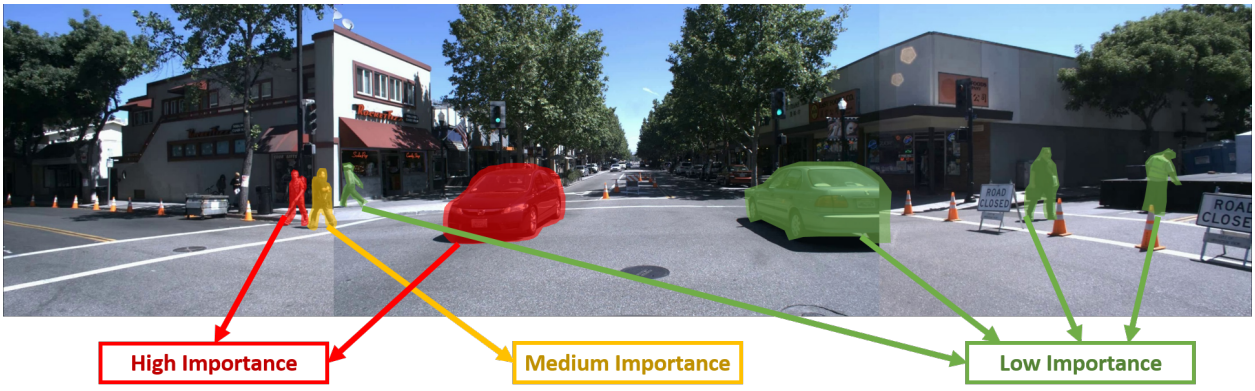


Figure 5.1: Illustration of an ideal road-user importance estimation during a left turn maneuver. In a driving scenario, there can be many road-users. However, when given an ego-vehicle's path, some road-users are more important for decision-making.

- Proposing a novel context aware architecture and a new way of representing the global context of the scene based on predicting the intention (future path) of the ego-vehicle.
- Systematic quantitative and qualitative evaluation of the proposed method against several baselines.

Recently many efforts have been devoted to development of vehicles with higher level of autonomy based on scene understanding.

For instance, driver's gaze has been wildly studied for determining saliency map and intention prediction relying only on fixation maps (Pugeault and Bowden, 2015). (Underwood et al., 2011) inspects the driver's attention specifically towards pedestrians and motorbikes, and exploits object saliency. In (Palazzi et al., 2017), a computer vision based model is proposed to predict saliency by conducting a data-driven study on drivers' gaze fixations. However, driver's gaze is not always a valid indication of saliency since the driver might look at many unimportant objects in the scene as well.

Different from our proposed method, prediction of important objects is also studied by (Kuen et al., 2016; Li et al., 2016). (Kuen et al., 2016) uses recurrent attention and convolutional-deconvolutional network to tackle the salient object detection problem. Furthermore, the proposed model in (Li et al., 2016) takes a strategy for encoding the underlying saliency prior information, and then sets up a multi-task learning scheme for exploring the intrinsic correlations between salient object detection and semantic image segmentation. However, these methods are not applicable to road-user importance estimation in driving scenario which highly depends on the ego car's intention and its interaction with other road-users.

Another approach for solving the saliency estimation problem in autonomous driving is using sensor-based methods. LiDAR (Halterman and Bruch, 2010), radars, lasers and sonars (Park et al., 2003) are popular sensors to detect surrounding objects in autonomous systems. For instance, (Sheu et al., 2007) uses smart antennas and proposes a distance awareness system for important object estimation. The model proposed by (Chen et al., 2017) combines the front view of the LiDAR point cloud with region-based features from the bird's eye view for 3D object detection. However, the salient objects are not necessarily

the nearest object (e.g. nearest object like a parked car may not pose as much a threat as a pedestrian intending to cross ego-vehicle's path further down the road). Therefore, visual information is essential for practical autonomous driving systems. For more details about the history of using different sensors and methods for autonomous driving systems please refer to (Janai et al., 2017).

Different from recent works based on estimating a general saliency map of the scene (i.e., a heat map which gives each pixel a relative value of its level of saliency), our proposed method is able to specifically estimate the importance level of all the road-users based on the scene context and ego car's intention. Furthermore, unlike the works based on estimating the driver's gaze fixation map (Cornia et al., 2018), in this research we propose a road-user importance estimation method based on human-centric importance annotation.

5.2 Method

An overview of the proposed model which we refer to as Context Aware road-user importance Estimation model or iCARE is shown in Figure 5.2. There are two main stages in the proposed model. First, an important road-user proposal generator provides potentially important road-user proposals and then in the second stage, context is incorporated into the system in order to estimate the importance level of road-users. The next subsections describe these stages in more details.

5.2.1 Important road-user Proposal Generation

In a busy intersection, there can be many road-users in the scene and our proposed model is designed to first select potentially important road-users among them. Thus, first we need a hard attention mechanism to pick these potentially important road-users out of the whole scene. In this stage, a detection model (Ren et al., 2015) is exploited in order to generate the potentially important road-users. The important road-user proposal generation is performed by using the Region Proposal Network (RPN) which predicts object proposals and at each spatial location, the network predicts a class-agnostic objectness score and a

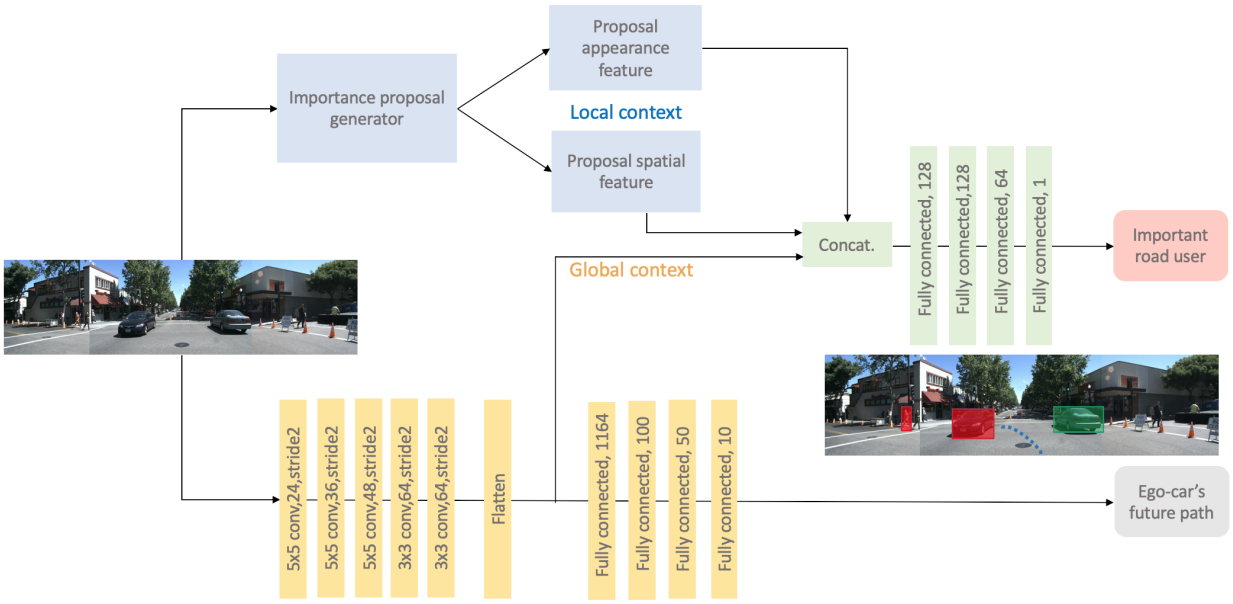


Figure 5.2: The iCARE model exploits local (i.e. appearance, location) of road-users and global (i.e. intention based context) of the scene to estimate importance of respective road-users.

bounding box refinement for anchor boxes of multiple scales and aspect ratios. Using non-maximum suppression with an intersection-over-union (IoU) threshold, the top box proposals are selected. Then, region of interest (RoI) pooling is used to extract a fixed-size feature map for each box proposal. These feature maps then go through the final fully connected layers where the class label (i.e, important road-user) and bounding box refinements for each box proposal are obtained. This stage of our model is applied to select more likely candidates of importance, where further consideration into context is necessary to accept or reject the proposal.

5.2.2 Context Aware Representation

iCARE takes advantage of the local (i.e. appearance, location) and global context towards estimating importance of road-users.

Local Appearance Feature

The appearance feature of the road-users contains very rich information about orientation, dynamics, intention, distance, etc. of the road-users. In this work, we use the Inception-ResNet-V2 (Szegedy et al., 2017) model as feature extractor in conjunction with

the road-user proposal generator model. To train the important road-user proposal generator model, we first initialize it with Faster R-CNN trained on COCO object detection dataset (Lin et al., 2014) and then train it based on the important road-user annotations in our data set. To generate the appearance feature for each potential important road-user proposal, we take the final output of the model and select all bounding boxes where the probability of belonging to the "important" class exceeds a confidence threshold. For each selected proposal, the appearance feature is defined as the output of the region of interest (RoI) pooling layer for that bounding box.

Location Feature

road-users with different sizes and distances to the ego-vehicle have different attributes which can make them better distinguishable. In our model, for each proposal of important road-user, we consider a $4D$ vector as the location feature f_{loc} which is defined as:

$$f_{loc} = [((x_{max} + x_{min})/2, y_{max}), h, w], \tag{5.1}$$

where $((x_{max} + x_{min})/2, y_{max})$ is the coordinate of the middle bottom point of each proposal bounding box and h and w are the height and width of the bounding box, respectively. The location feature helps the system to learn the correlation between proximity, mass and importance.

Intention-based (Global) Context

Intention of the ego car plays a major role in estimating the importance of road-users. For instance, if the ego car's intention is to make a left turn in an intersection, then road-users on the right side of the intersection may be considered relatively less important road-users. In order to incorporate the intention of the ego car, we design a model (inspired by (Bojarski et al., 2016), where the model learns a mapping between an image and instantaneous steering angle) which takes as input a single image of the scene and predicts a $10D$ vector of the future path of the car. The future path vector is constructed of 10 steering angle values, representing the next 10 spatial steps of the car with 1-meter equal spacing. The model for predicting the future path (shown in blue in Figure 5.2) consists of convolution layers

followed by fully connected layers with batch normalization and drop out layers in between. The flatten feature of the last convolution layer is used as the context feature.

Feature Fusion

In this part of the model, the local (i.e., appearance and location features) and global features (i.e., intention-based context) are concatenated together and followed by 4 fully connected layers to estimate importance of a road-user (shown in magenta in Figure 5.2). To study the effect of intention and context, we consider a version of our model were instead of context, the $10D$ ground truth future path vector is used as an input to the model and the results are compared. Furthermore, an ablation study is performed on combinations of the features (i.e. appearance, spatial, intention context and the future path of the car as input) which will be elaborated in the next section.

5.3 Experiments

5.3.1 Data set

The data set used for training and evaluation of the proposed iCARE model is collected at Honda Research Institute and includes the aligned-view images that are generated by putting together (and aligning) the images from three cameras (i.e., left, center and right views). The data set consists of 6 hours of driving including around 2.7 hours of intersections. The data is collected from driving on the streets of Mountain View and Sunnyvale in California. There are 743 total intersection segments which include 307908 total frames.

The important road-users in image sequences are annotated every 30 frames (i.e., 1 annotation per second). The annotations are generated by human annotators who are given the video sequences and instructions about the driving rules. There are 9995 total frames of annotations, of which 6924 total road-users are annotated as important. An annotated image may include between zero to five important road-users. Some examples from the data set are illustrated in Figure 5.3.

Figure 5.3: Examples of the images and annotations in our data set

The data is split to training set and testing set with 13624 and 4749 images, respectively. Only images with annotation (i.e., with at least one important road-user in them) are used for training, but testing is performed on all images in the test set.

5.3.2 Implementation Details

In our model, the aligned view images are re-sized from their original size (i.e., 4394×1100) to 1024×275 before going into the deep neural network model. Tensorflow v1.4 is used as our deep learning framework on a Tesla V100-SXM2 NVIDIA GPU with 32 gigabytes of memory.

For the intention-based context extractor branch, we use 5 convolution layers followed by 4 fully connected layers. Batch normalization (Ioffe and Szegedy, 2015) is used for faster training and also there are drop out layers (with keep-prob $= 0.6$) between the fully connected layers to avoid over fitting. The intention-based context branch is trained and tested based on the same data split used for important road-user proposal generation. The last convolution layer of this model is extracted and flattened to $1164D$ vector and used as the context feature

for each image. Mean Square Error (MSE) is used as loss function for training of this part of the model.

The local features from the proposal generator and the context features are fused (i.e., concatenated) and then go through 4 fully connected layers with 128, 128, 64 and 1 neurons with batch normalization and drop out between the layers. Binary cross entropy is used as the loss function for the final classification step (i.e., important vs not important road-users). Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and learning rate of 0.01 is used for optimization of the loss functions in all parts of our model. Moreover, the Relu (Nair and Hinton, 2010) is used as non-linearity throughout our model.

5.3.3 Evaluation and Results

Since the number of not important road-users is larger than number of important road-users in our data set, we need to deal with data imbalance problem in training our model. In fact, in the training set there are 4699 important samples and 8925 not important samples. In order to solve the data imbalance problem we assign appropriate weights for the loss terms of each class (i.e., $1 : 2$). Furthermore, data imbalance causes the classification accuracy metric (i.e., unweighted accuracy) to not be able to precisely estimate the performance of the model. Hence, we use the precision-recall curve and $F1$ score to evaluate our model. The $F1$ score is the harmonic average of the precision and recall, and it reaches its best value at 1 (perfect precision and recall) and worst at 0.

The precision-recall curves for different experiment settings are shown in Figure 5.4. The $F1$ score is shown in parenthesis for each experiment, as well. In Figure 5.4, the red curve (denoted as appearance (0.65)) is the setting where only the appearance feature taken from the important road-user proposal is used to estimate the road-user importance. Using only the appearance features, our model achieves $F1$ score of 0.65.

The green curve (denoted as: appearance + spatial + path (0.67) in Figure 5.4) corresponds to the setting where the future path is used as an input to the model and is concatenated with appearance and location features. This combination of the features yields an $F1$ score of 0.67 which is 3% higher than using only the appearance feature. The light blue curve which achieves the best performance ($F1 = 0.69$) is when the intention-based context

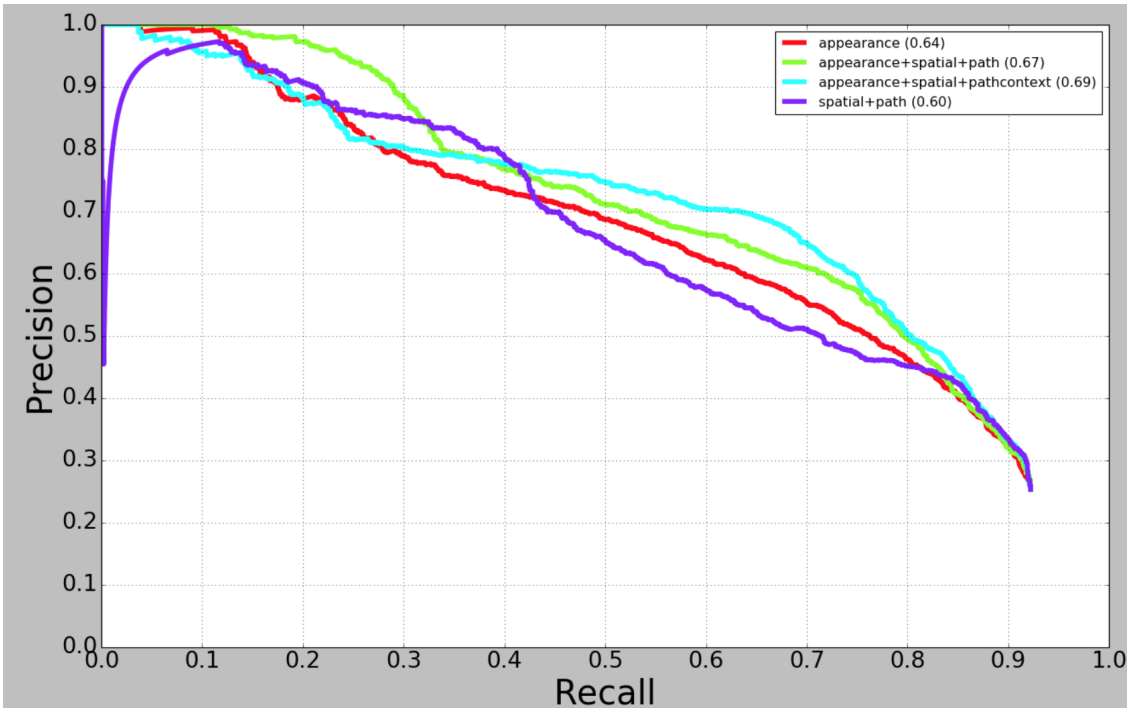


Figure 5.4: Precision-recall curves (and $F1$ scores) for different experiment settings. Best viewed in color.

of the scene is incorporated to the system along with the local features (i.e., appearance and location features). This result shows the importance of exploiting intention-based global context representation of the scene.

Moreover, the magenta curve illustrates the experiment setting where the appearance of the road-users is not exploited but only location and future path (as input) are used. The results show that this setting achieves the lowest performance of $F1 = 0.60$ as it is expected. It shows the importance of using the appearance features which has rich information about the orientation, type, etc. of the road-users. It is worth noting that the reported results in this section are all based on testing on all the images in the testing set where event images without any ground truth important road-user annotation are also considered for testing. Testing only on images with annotation achieves results with around 10% improvement compared to the reported results. The qualitative results of road-user importance estimation using our proposed model are illustrated in Figure 5.5.

In this Figure, the blue bounding boxes correspond to the ground truth annotation and the yellow bounding boxes correspond to the estimation of the model when using only

the appearance feature. The red bounding boxes show the result of our iCARE model which considers the intention based context of the scene as well as the appearance and location features of the road-users. It can be seen in Figure 5.5-(a), when using only the appearance feature, model gives false positive estimations for the cyclist on the right hand side of the intersection. Also, for the car on the most left hand side, only the iCARE can successfully detect the important road-users based on the learned intention from the scene context. Furthermore, Figure 5.5-(b) and (c) show two very common cases in our test set that incorporating the ego car's intention based context helps to get rid of the false positive estimations when using only the appearance of the road-users. In one experiment we also compare the performance of our model when using the future path of the ego car as input to the model versus when the model uses the context from path prediction (i.e., iCARE). Figure 5.6 shows some examples of this experiment.

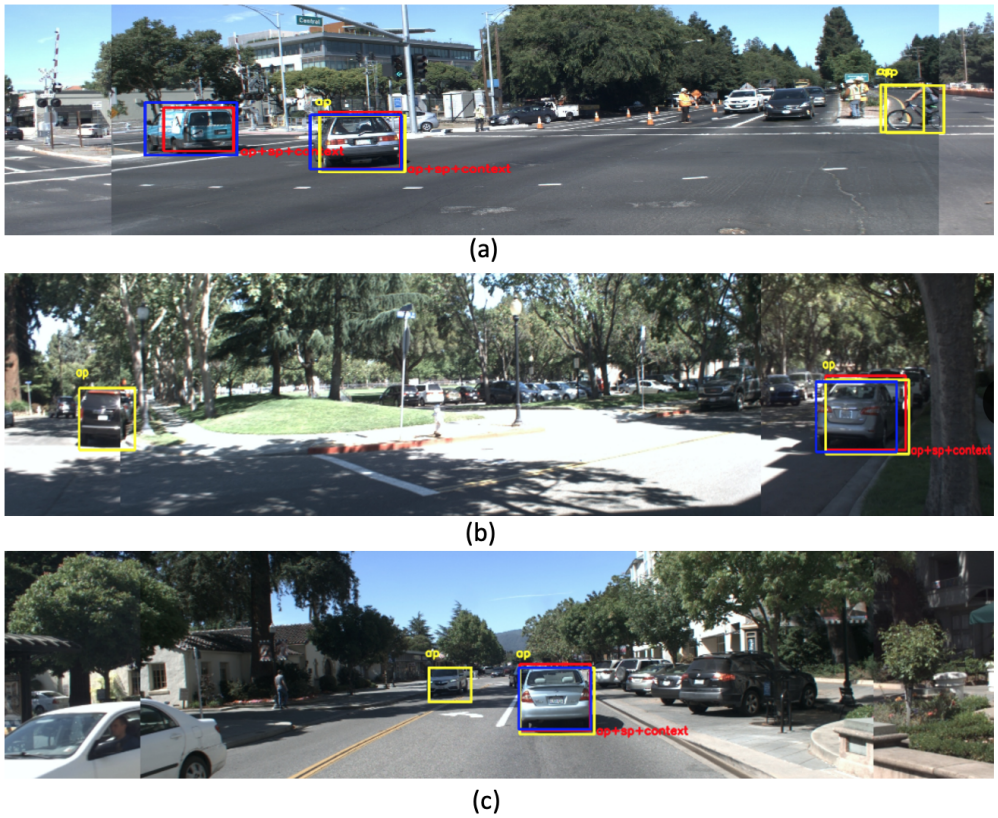


Figure 5.5: Examples of performance comparison of using appearance feature only (yellow) vs iCARE (red) and ground truth (blue).

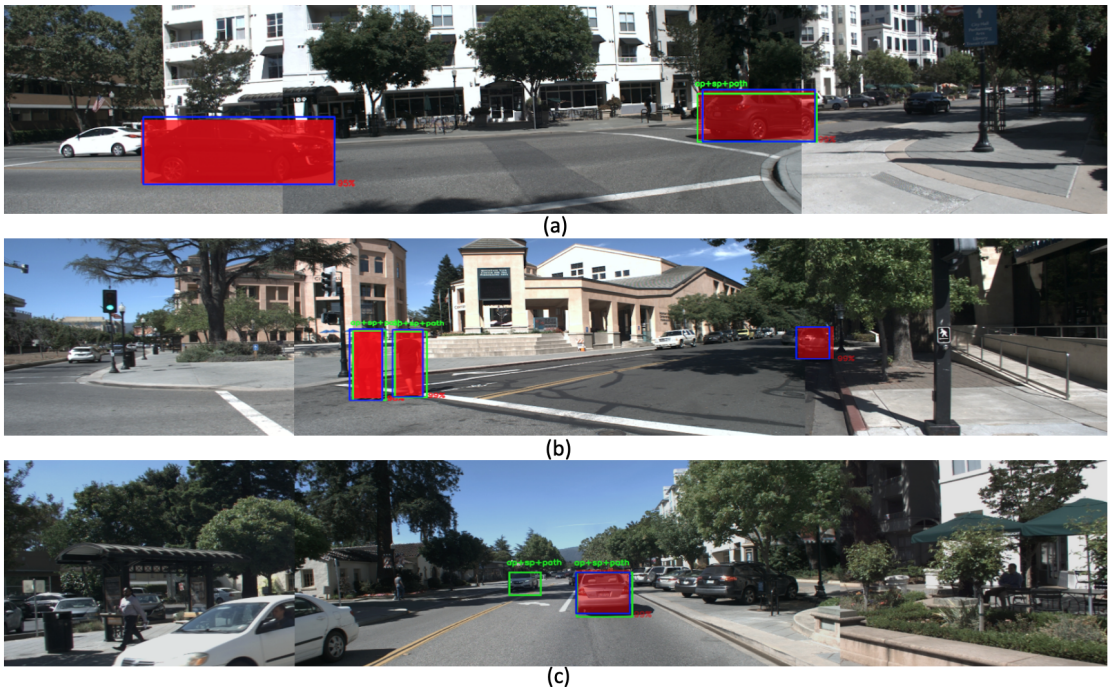


Figure 5.6: Comparison of performance of iCARE model (red) vs fusion of appearance, spatial and input future path features (green). The blue bounding boxes show the ground truth annotations for important road-users. The intensity of the red color shows the level of importance of each road-user estimated by iCARE. Best viewed in color.

In this Figure the red layover color demonstrates the output of the iCARE model and the intensity of the red color illustrates how important that road-user is based on the prediction of our model. The blue bounding box corresponds to the ground truth and the green bounding box corresponds to the experiment setting where the future path of the ego car has been used as input feature (along with the appearance and location feature). It can be observed that using the $10D$ future path vector as input is not as effective as using the future path context. For instance, it can be observed in Figure 5.6-(a)-(b) that when using the future path as input, model can not estimate the important road-users properly, and also it sometimes leads to false positives as it is shown in Figure 5.6-(c).

Moreover, the estimation error in predicting the future path of the car is shown in Figure 5.7. It can be observed that the estimation error increases as the distance to the ego vehicle increases.

Some examples of our model's failure estimations are shown in Figure 5.8. For instance in Figure 5.8-(a), the iCARE model is not able to detect the left car in the intersection. This is

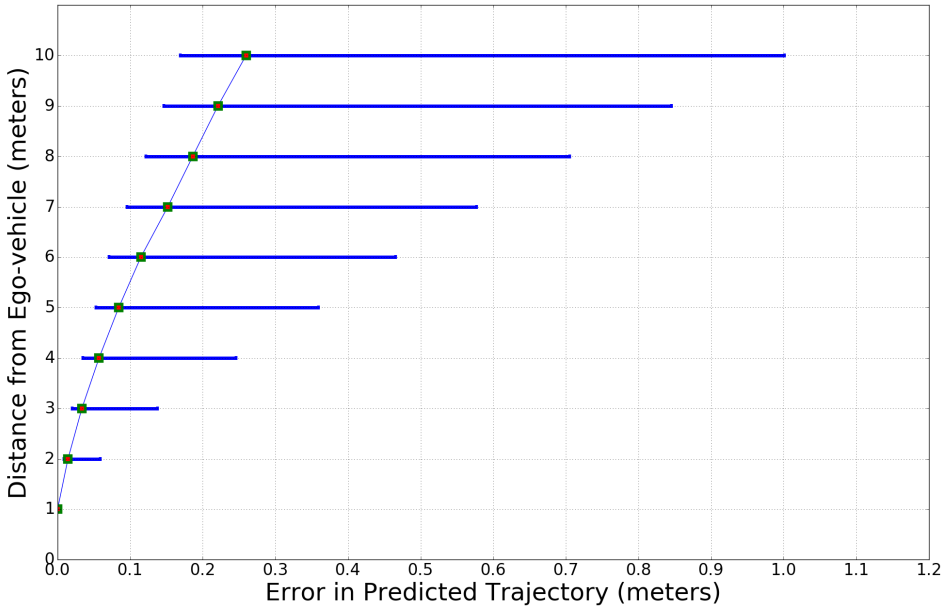


Figure 5.7: Ego-vehicle future path prediction error versus distance from ego-vehicle.

mainly due to lack of road-user's intention information (i.e. if the white car's intention is to turn right then it is not indeed important, but if it is going straight it should be considered as important). Moreover, when some road-users are very far away from the ego car (Figure 5.8-(b)), iCARE estimates them as not important. Another failure case is due to mis-detection and other unavoidable causes (Figure 5.8-(c)). Interestingly, sometimes estimations of the iCARE does make sense even though those road-users have not been annotated as important (e.g., traffic sign in Figure 5.8-(c)).

In another experiment, we investigate the subjectivity issue in road-user importance estimation. In fact, even though most drivers agree on the obvious important road-users (e.g., a pedestrian in front of the moving ego car, etc.), different drivers might have different opinion about importance of some of the road-users. In order to study this subjectivity, annotation from a different annotator is used to test our model. The performance of iCARE versus appearance-based baseline when trained with first annotation and tested with second annotation is shown in Figure 5.9. This Figure shows precision-recall curves of the proposed iCARE model (shown in light blue) with $F1 = 0.59$ and appearance-based baseline (shown in red) with $F1 = 0.53$. It can be observed that even though the iCARE model achieves

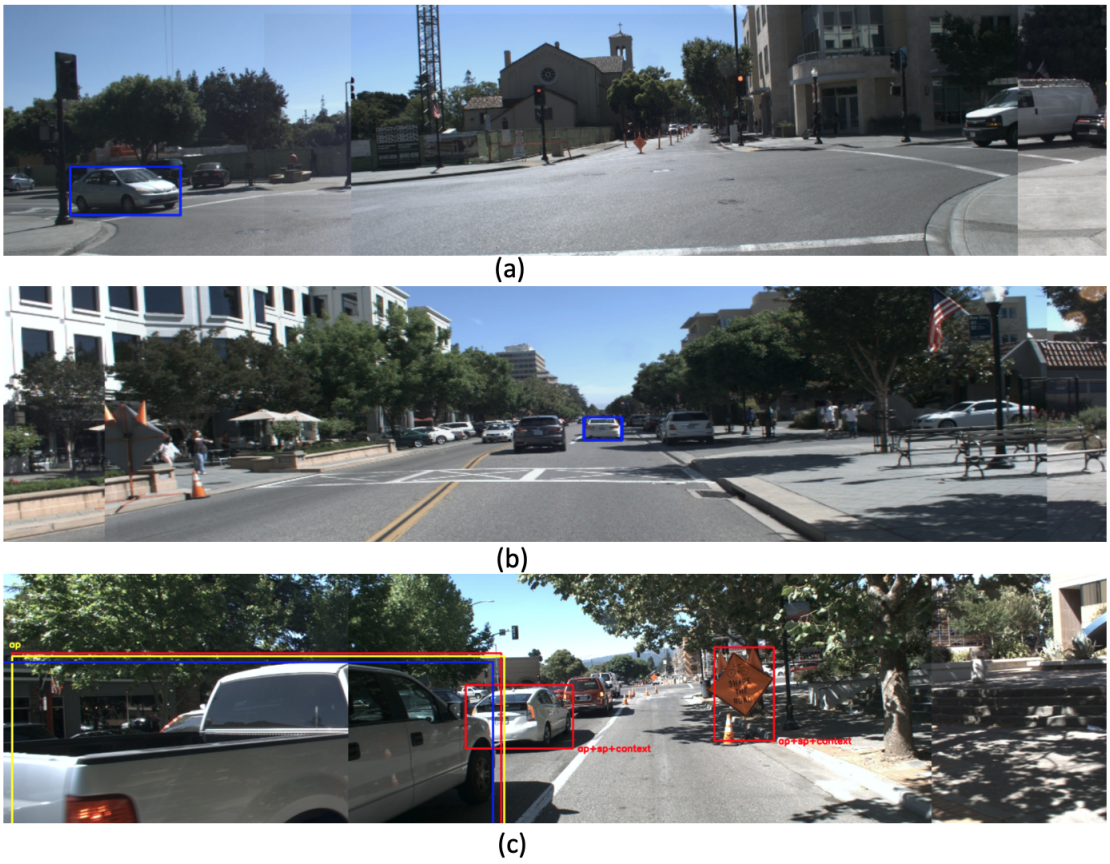


Figure 5.8: Three examples of failure cases of iCARE model estimations. (red: iCARE estimation, blue: ground truth, yellow: using only appearance feature for importance estimation.)

lower accuracy (compared to train and test with same annotation), it still works fairly well and has a consistent behavior with our previous results.

5.4 Conclusion and Future works

In this research, we investigated the effect of ego car's intention and its context on estimating road-users importance using only images taken from 3 cameras in front of the car. The proposed iCARE model estimates the important road-users based on a 2-stage recognition framework, where the first stage generates important road-user proposals using an importance-guided training scheme. In the second stage, model selectively picks the most important road-user proposals by taking into account the location and intention context information. Our future work is to incorporate the intention of the road-users into our model

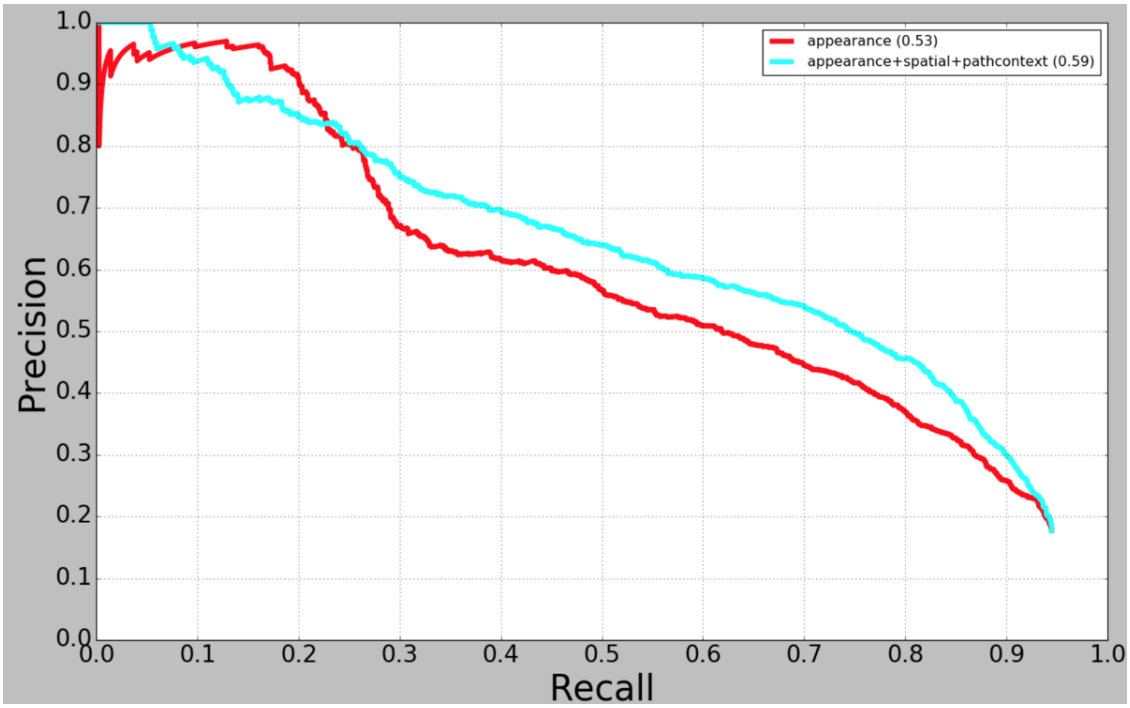


Figure 5.9: Precision-recall curves for iCARE (light blue) and the baseline (red) when trained on the first annotation and test on the second annotation.

which plays an important role in determining which road-user is important. Furthermore, incorporating other contextual information (e.g., depth, motion, etc.) can be an interesting line of future research for road-user importance estimation.

Chapter 6

Class-Discriminative Meta-Learning based Few-Shot Learning

Although deep learning-based approaches have been very effective in solving problems with plenty of labeled data, they suffer in tackling problems for which labeled data are scarce. In few-shot classification, the objective is to train a classifier from only a handful of labeled examples. In this research, we propose an attention-based context-aware query embedding encoder for incorporating support set context into query embedding and generating more discriminative and task-dependent query embeddings.

Moreover, we propose a few-shot learning framework based on structured margin loss which takes into account the global structure of the support set in order to generate a highly discriminative feature space where the features from distinct classes are well separated in clusters. The task-dependent features help the meta-learner to learn a distribution over tasks more effectively. Extensive experiments based on few-shot, zero-shot and semi-supervised learning on three benchmarks show the advantages of the proposed model compared to state-of-the-art.

6.1 Introduction

Deep learning has made major advances in many areas, but still has limitations when it comes to problems with limited number of labeled data. In practice, many learning problems require

rapid inference from small amount of data. In particular, many practical recognition systems should be able to recognize a new category from a handful of training images. Humans on the other hand are able to rapidly learn new classes. For example, a child can learn to recognize a new object by only seeing one picture of that object. Human can recognize objects even without seeing the examples of that object category and just by hearing the description of that object (similar to zero-shot learning). This significant gap between human and machine learning provides fertile ground for few-shot learning developments.

Few-shot classification is a task in which a classifier must be able to generalize from few examples. Recently there has been a surge of interest in using meta-learning (learning-to-learn) for few-shot learning (Snell et al., 2017; Vinyals et al., 2016; Ren et al., 2018; Sung et al., 2018). These approaches use a meta-learning strategy which includes extracting some transferable knowledge from a set of tasks and transferring the knowledge to quickly adapt to new tasks without suffering from the overfitting that might happen when applying deep models to problems with small amount of data. Specifically, these meta-learning based models utilize sampled mini-batches called episodes during training, where each episode is designed to mimic the few-shot task by sub-sampling classes as well as data points. The use of episodes makes the training problem more faithful to the test environment and thereby improves generalization (Vinyals et al., 2016). In fact, the meta-learner learns a strategy for generalizing to an unseen task from a similar task distribution. Here instead of learning the distribution of data samples (as in regular machine learning algorithms), the model learns the distributions of tasks.

Several successful directions have been explored recently for meta-learning-based few-shot learning, including learn to fine-tune (Finn et al., 2017; Ravi and Larochelle, 2016), sequence based methods (Santoro et al., 2016), and metric learning models (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018). However, there are still challenges in solving the few-shot learning problem. For instance, even though reducing the intra-class variation is a very critical factor in the current few-shot classification problem setting, recent works seldom explicitly study it. In this work, we address this issue by defining a structured-based margin loss to explicitly decrease the intra-class distance between feature embedding of each class in the support set and create a *structured support set embedding*. The structured-based margin

considers the relationship between all the support set samples in minimizing the loss and guide the model to learn a deep metric to cluster the support set embeddings and generates a highly discriminative feature space where all classes are well separated. We refer to the proposed Class-Discriminative Few-Shot learning framework in this research as CDFS.

In episode-based few-shot learning frameworks a task is defined based on context of support set and its relationship with the query in each episode. It has been shown in (Oreshkin et al., 2018), that incorporating the task information to the feature embedding can highly improve the performance of few-shot classification as in Prototypical Networks (Snell et al., 2017). The proposed context-aware query embedding in this research incorporates the task information into query embedding in each episode using attention mechanism and 1-D CNN.

Besides few-shot learning, we also show performance of our proposed model for zero-shot classification. In the zero-shot setting, each class comes with a category description (meta-data) giving a high-level description of the class rather than a small number of labeled examples. We therefore learn an embedding of the meta-data into a shared space to serve as the prototype for each class. Classification is performed, as in the one-shot scenario, by finding the nearest class prototype for an embedded query point.

The main contributions of this research are summarized as follows:

- Regularizing the few-shot classification setting with a structured-based margin loss which takes into account the global structure of the support set feature space and learns to explicitly reduce the intra-class variation in order to map the data to a highly discriminative feature space where the few-shot classification is most effective.

- Proposing a context-aware query embedding module which takes into account the support set's context and generates task-dependent feature representations which would help the meta-learner to learn a distribution over tasks more effectively.

- Performing extensive experiments based on few-shot, one-shot, zero-shot and semi-supervised learning schemes to show the advantages of the proposed model compared to state-of-the-art.

Recently there has been a resurgence of interest in few-shot learning based on meta-learning (Finn et al., 2017; Vinyals et al., 2016; Snell et al., 2017; Ravi and Larochelle, 2016; Santoro et al., 2016; Munkhdalai and Yu, 2017; Sung et al., 2018). The existing meta-learning models for few-shot classification can be divided into three types: the learning to fine-tune based, RNN based, and metric learning based. For instance, in (Finn et al., 2017) the MAML model aims to meta-learn an initial condition that is good for fine-tuning on few-shot problems. The model in (Ravi and Larochelle, 2016) is an LSTM-based optimizer that is trained to be specifically effective for fine-tuning. In (Santoro et al., 2016), a recurrent neural network iterates over examples of given problem and accumulates the knowledge required to solve that problem in its hidden activations. However, these recent works either require fine-tuning the target problem (Finn et al., 2017; Ravi and Larochelle, 2016), or need the use of complex recurrent neural network (RNN) architectures (Santoro et al., 2016; Vinyals et al., 2016), or are based on complicated inference steps (Fei-Fei et al., 2006). In our work, the model is simple and fast and does not need any additional process such as fine tuning. Moreover, we avoid the complexity of recurrent networks, and the issues involved in ensuring the adequacy of their memory. Instead our proposed approach is defined entirely with feed forward convolution neural networks.

The metric based few-shot learning has attracted a lot of interests recently (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018). The basic idea is to learn a metric which can map similar samples close and dissimilar ones distant in the metric space so that a query can be easily classified. Various metric based methods such as siamese networks (Chopra et al., 2005), matching networks (Vinyals et al., 2016), prototypical networks (Snell et al., 2017), and relation networks (Sung et al., 2018) have been proposed. They differ in their ways of learning the metric. For instance, very recently the relation network (Sung et al., 2018) proposed to replace the fixed metric learning part (e.g., Euclidean distance) of the previous works with a deep metric for comparing the relation between images.

The success of metric based methods relies on learning a discriminative metric space. The proposed method in this research can be categorized as the metric learning based framework and the global version of the triplet loss (Schroff et al., 2015). To reach the full potential of metric based few-shot learning, we augment the classification loss with a structure-based

deep metric learning regularization which enforces the model to map the samples in the support set to well separated clusters in the embedding space. This regularization is based on an improved version of deep metric learning framework in (Oh Song et al., 2017) with no need of sample selection and greedy algorithm. Unlike the metric learning methods based on contrastive (Chopra et al., 2005) or triplet (Schroff et al., 2015) loss that are defined in terms of data pairs or triplets, our approach takes into account the global structure of the embedding space. In fact, the structured margin term in the loss function measures the quality of clustering the data by taking into account the relationship between all the data points in the mini batch at once (instead of data pairs or triplets). Furthermore, this deep learning based metric learning framework does not require the training data to be preprocessed in rigid paired or triplet format and uses a structured prediction framework (Tsochantaridis et al., 2004; Joachims et al., 2009) to ensure that the score of the ground truth clustering assignment is higher than the score of any other clustering assignment.

Taking advantage of contextual information in the support set is critical in episode-based few-shot learning models. A framework for context modeling in the support set was proposed in (Vinyals et al., 2016) based on a bi-directional LSTM. However, as the number of classes and shots increases, the model is required to learn longer and more complex dependencies, which negatively affects both generalization and efficiency. Furthermore, it imposes an arbitrary ordering on the support set by using bi-directional LSTM (i.e., the embedding changes if we shuffle the support set samples). Moreover, the meta-learner architecture proposed in (Mishra et al., 2017) combines temporal convolutions (which aggregate contextual information from past) with causal attention which pinpoints to specific pieces of information. In this research, we propose a simpler but effective context-aware query embedding framework based on attention mechanism and 1-D CNN for taking into account the context of the support set and its relationship (i.e., task) with query embedding. The proposed query encoder makes the query embedding task-dependent which helps learning a meta-learner with higher generalization power.

6.2 Method

In this section we first describe the meta-learning based few-shot classification. We then elaborate on components of our proposed model including structured support set embedding and context-aware query embedding modules.

6.2.1 Few-Shot Classification

The meta-learning based few-shot classification is defined based on episodic training. The idea behind the episodic paradigm is to simulate the few-shot task that will be encountered at test time. In each training iteration, an episode is formed by randomly selecting N_C classes from the training set with K labeled samples from each class to act as the support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $m = K \times N_C$ and a query set $\mathcal{Q} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_Q}$ of different examples from the same N_C classes. Each $\mathbf{x}_i \in \mathbb{R}^D$ is an input vector of dimension D and $y_i \in \{1, 2, \ldots, N_C\}$ is a class label. Training on such episodes is done by feeding the support set $\mathcal{S}$ to the model and updating the model's parameters to minimize the loss of its predictions for the examples in the query set $\mathcal{Q}$. This form of training allows the model to extract transferable knowledge based on different classification tasks seen in the episodes so the model can exploit this knowledge in testing stage to classify the query samples coming from new unseen classes.

In the proposed model, we employ a few-shot learning structure based on episodic training as in Prototypical Networks (Snell et al., 2017) which uses the support set $\mathcal{S}$ to extract a prototype $\mathbf{c}_j \in \mathbb{R}^N$ from each class $j = 1, \ldots, N_C$ through an embedding function $f_\phi(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^N$, where ϕ is the learnable parameters of the neural network. Each prototype is defined as the mean vector of the embedded support points belonging to its class:

$$\mathbf{c}_j = \frac{1}{|\mathcal{S}_j|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_j} f_\phi(\mathbf{x}_i), \tag{6.1}$$

where $i = 1, \ldots, K$. The samples in the query set are then classified based on their distance to the prototype of each class and a distribution over classes for a query point $\mathbf{x}_q$ is defined based on a softmax over distances to the prototypes in the embedding space (Snell et al.,

2017):

$$p_\phi(y = j|\mathbf{x}_q) = \frac{\exp(-d(\mathbf{c}_j, f_\phi(\mathbf{x}_q)))}{\sum_{j'} \exp(-d(\mathbf{c}_{j'}, f_\phi(\mathbf{x}_q)))} \tag{6.2}$$

It has been shown in (Snell et al., 2017), that the prototype in Eq. 6.1 yields cluster representatives with the prototype as the cluster center and there is one cluster per class when a Bregman divergence such as squared Euclidean distance is used.

In order to learn more discriminative embeddings for the few-shot learning task, in this research we propose to impose a constraint based on structured margin on support set, to explicitly enforce the class separation in embedding space based on the global structure of the support set (Sec. 6.2.2). Furthermore, a context-aware query embedding module is proposed to create task-dependent query features and also to pull the feature embedding of the query towards corresponding class prototype in support set (Sec. 6.2.3). A toy example showing the effect of proposed Class-Discriminative Few-Shot learning (CDFS) model on feature space in a 5-shot, 3-way classification task is demonstrated in Figure 6.1. The model architecture is shown in Figure 6.2.

6.2.2 Structured Support Set Embedding

Similar to metric-based few-shot learning methods (Snell et al., 2017; Vinyals et al., 2016; Sung et al., 2018), our model learns a nonlinear embedding function $f_\phi(\mathbf{x})$, parameterized

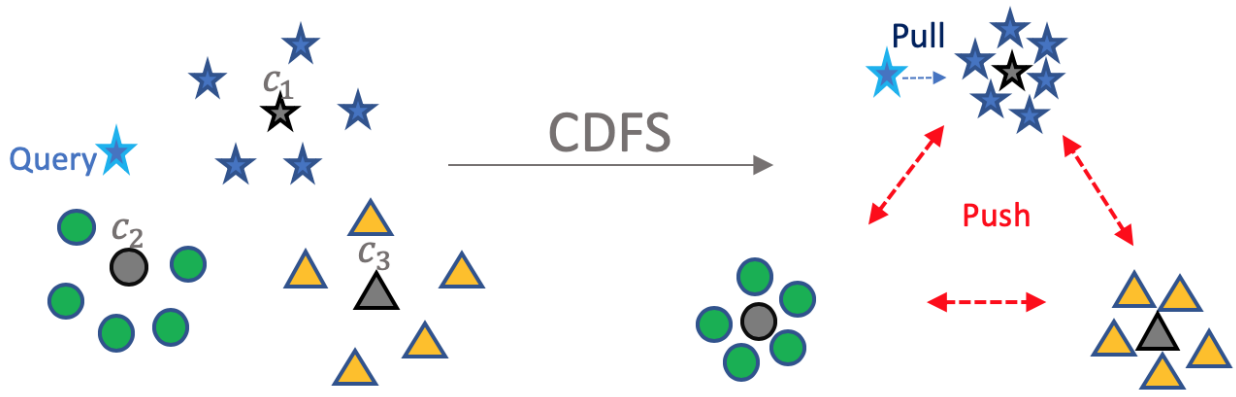


Figure 6.1: Toy example showing the effect of proposed Class-Discriminative Few-Shot learning (CDFS) model on embedding space.

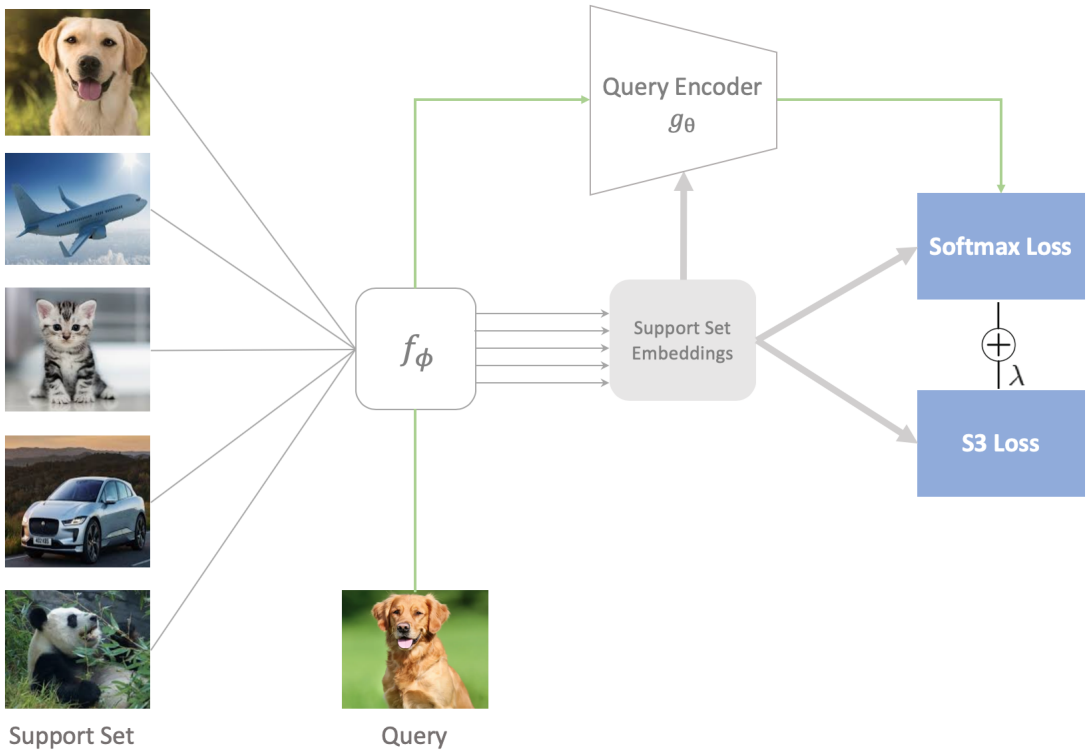


Figure 6.2: Model architecture for 5-way, 1-shot classification.

as a neural network, that maps examples into a space where examples from the same class are close and those from different classes are far apart. The embedded point $f_\phi(\mathbf{x})$ is then classified by a classifier, e.g., the softmax classifier. In this research, our objective is to learn highly discriminative features with the joint supervision of softmax loss and Structured Support Set (S3) loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{softmax}} + \lambda \times \mathcal{L}_{S3}, \tag{6.3}$$

where $\mathcal{L}_{\text{softmax}}$ is:

$$\begin{aligned} \mathcal{L}_{\text{softmax}} = \frac{1}{N_Q} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{Q}_j} \Big[& d(\mathbf{c}_j, g_\theta(f_\phi(\mathbf{x}_i))) + \\ & \log \sum_{j'} \exp(-d(\mathbf{c}_{j'}, g_\theta(f_\phi(\mathbf{x}_i)))) \Big], \end{aligned} \tag{6.4}$$

which is simply defined based on the average negative log-probability of the correct class assignments, for all query examples. $g_\theta(.)$ is the context-aware query embedding function

to be described in Sec. 6.2.3 and λ is a scalar used for balancing the two loss functions. $\mathcal{L}_{S3}$ is the Structured Support Set (S3) loss which guides the training by enforcing a margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$ based on the global structure of the support set as follows:

$$\mathcal{L}_{S3}(X, f_\phi) = \left[F(X, \hat{\mathbf{y}}; f_\phi) + \gamma \Delta(\mathbf{y}, \hat{\mathbf{y}}) - F(X, \mathbf{y}; f_\phi) \right]_+, \tag{6.5}$$

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \text{AMI}(\mathbf{y}, \hat{\mathbf{y}}), \tag{6.6}$$

where $[z]_+ = \max(z, 0)$ and $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is the set of samples in the support set and $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the predicted and ground-truth support set labeling assignments, respectively. This loss encourages the model to learn an embedding function f_ϕ such that the ground truth labeling score for the support set $F(X, \mathbf{y}; f_\phi)$ is greater than the score for any other label assignments of the set $F(X, \hat{\mathbf{y}}; f_\phi)$, at least by the structured margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$. F is defined as a scoring function that encourages the embeddings of samples in each class to be as close as possible to the prototype of that class and reduces the intra-class distance between embeddings of each class and results in a compact feature representation of that class around its prototype as follows:

$$F(X, \hat{\mathbf{y}}; f_\phi) = -\sum_{\mathbf{x}_i \in X} \min_j ||f_\phi(\mathbf{x}_i) - \hat{\mathbf{c}}_j||_2^2, \tag{6.7}$$

where $\mathbf{x}_i$ is the ith data sample (e.g., image) in the support set and $j = 1, \ldots, N_C$.

The structured margin has been used in structure prediction problems such as structured SVM (Finley and Joachims, 2008), structured KNN (Pugelj and Džeroski, 2011), etc., where the problem involves predicting structured objects. In our problem the structured output is defined as the labeling configuration of the support set. We define the structured margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$ to measure the quality of the label assignment of the support set as in Eq. 6.6, where AMI is the Adjusted Mutual Information and is defined as:

$$\text{AMI}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{MI(\mathbf{y}, \hat{\mathbf{y}}) - E\{MI(\mathbf{y}, \hat{\mathbf{y}})\}}{\max\{(H(\mathbf{y}), H(\hat{\mathbf{y}}))\} - E\{MI(\mathbf{y}, \hat{\mathbf{y}})\}}, \tag{6.8}$$

where MI is the mutual information which is a non-negative quantity which quantifies the information shared by the two label sets (i.e., clusterings), $E\{MI(\mathbf{y}, \hat{\mathbf{y}})\}$ is the expected value of the MI, and H is the entropy. The AMI takes a value of 1 when the two sets are identical and 0 when the MI between two sets equals the value expected due to chance alone. In fact, AMI is an adjustment of the MI score to account for chance. Our experiments show that using AMI gives slightly better results compared to NMI and other similarity measures. For more details about AMI please refer to (Romano et al., 2016).

Training of the model is performed by minimizing the average loss, iterating over training episodes and performing a gradient descent update for each. For more details about optimization for structured prediction please refer to (Tschiatschek et al., 2014; Lin and Bilmes, 2012; Oh Song et al., 2017; Tschiatschek et al., 2014). All parameters of our model lie in the embedding function and by using the combination of softmax loss and structured margin loss, the model learns a discriminative embedding function with two key learning objectives, inter-class dispersion and intra-class compactness as much as possible, which are essential to few-shot learning.

6.2.3 Context-Aware Query Embedding

The goal of this part of the model is to create task-dependent query embeddings and pull them towards their class prototypes based on the task context in each episode. Task-dependent query features help the meta-learner to learn a more effective distribution over the tasks. Let $f_\phi(\mathbf{x}_q)$ be the embedding of a query image taken from the CNN and $\mathbf{c}_j$ be the prototype of the jth class. For each sample in the query set $\mathcal{Q}$, a context vector $\mathbf{v}_q$ is extracted from the support set based on the similarity of the query embedding and the prototypes in the support set. The context vector is calculated using a content-based attention mechanism as follows:

$$a(\mathbf{c}_j, f_\phi(\mathbf{x}_q)) = \frac{\exp(-d(\mathbf{c}_j, f_\phi(\mathbf{x}_q)))}{\sum_{n=1}^{N_C} \exp(-d(\mathbf{c}_n, f_\phi(\mathbf{x}_q)))}, \tag{6.9}$$

$$\mathbf{v}_q = \sum_{j=1}^{N_C} a(\mathbf{c}_j, f_\phi(\mathbf{x}_q))\mathbf{c}_j, \tag{6.10}$$

where a represents the attention weight and d is, again, the Euclidean distance. The more similar a query embedding to a prototype of a class, the larger is the attention weight of that prototype in context vector. The content-based attention has the property that the context vector $\mathbf{v}_q$ will not be sensitive to the order of the prototypes in the support set since it is the weighted sum of them. In other words, the similarity information retrieved from the support set would not change if we randomly shuffle the prototypes in the support set. After calculating the context vector for each query member, $f_\phi(\mathbf{x}_q)$ and $\mathbf{v}_q$ get concatenated and go through a 1-D convolutional block. The convolutional block consists of batch normalization, ReLU activations and pooling. The output of the query encoder is $g_\theta(f_\phi(\mathbf{x}_q), \mathbf{v}_q)$ where θ is the trainable parameter of the encoder (i.e., 1-D CNN). Figure 6.3 illustrates the details of query embedding module by a toy example of 5-way classification task. The non-linear function g_θ is trained to infer the relationship between query and support set and modify the query feature to increase the discrimination power of the model. Figure 6.3 shows an example of how query encoder can modify the query embedding to be closer to the prototype of the matched class in the support set.

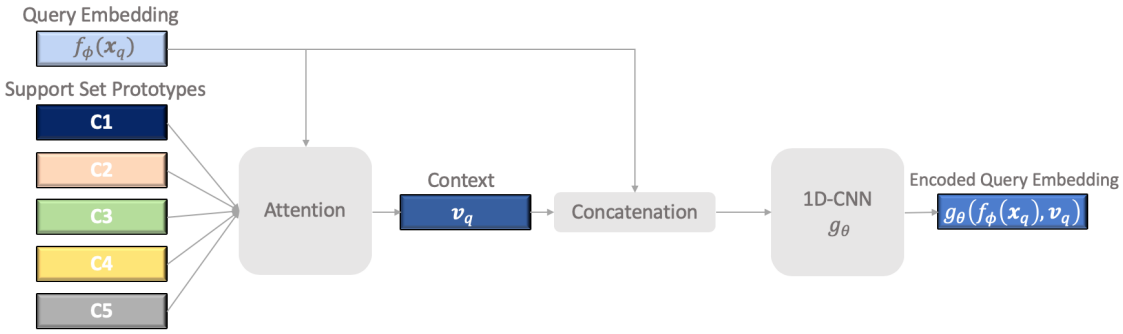


Figure 6.3: Context-aware query embedding architecture. In this example the query embedding $f_\phi(\mathbf{x}_q)$ and the top prototype $\mathbf{c}_1$ in the support set are in class 1. Change of blue color of the query embedding shows how the encoder pulls this feature towards the prototype of class one (i.e., $\mathbf{c}_1$) by incorporating the task context in episodes. In general, during training, the non-linear function g_θ learns how to modify the query embedding based on support set context to achieve optimum classification performance.

Table 6.1: Number of samples in episodes in different few-shot classification setting for Omniglot dataset during training.

Experiment	num. of queries	num. of support set samples	num. episode samples
5-way 1-shot	19	5	100
5-way 5-shot	15	25	100
20-way 1-shot	10	20	220
20-way 5-shot	5	100	200

6.2.4 Zero-Shot Learning and Semi-Supervised Adaptation

In Zero-Shot Learning (ZSL) we are given a class attribute vector $\mathbf{r}_j$ for each class instead of the support set of training data in the few-shot learning setting. In order to have our proposed model to work in zero-shot setting we define the prototype $\mathbf{c}_j = f_{\phi_2}(\mathbf{r}_j)$ to be the embedding of the attribute vector (different from the query embedding f_ϕ), since its modality is different from query images. Classification is performed, as in the few-shot scenario, by finding the nearest class prototype for an embedded query point.

Another capability of the proposed model in this research is to adapt to semi-supervised classification in testing stage. Specifically, in the semi-supervised scenario, the model needs to adapt to tasks which contain both labeled and unlabeled samples. We assume that we have access to a few labeled examples and many unlabeled examples from the classes in the support set. Since our model is able to generate highly distinguishable feature embeddings in form of separate clusters, unlabeled samples are clustered to the corresponding classes in test time. The prototypes are estimated at test time using the labeled and unlabeled samples and then the query samples are classified based on the nearest prototype. We will show that taking advantage of unlabeled samples can improve the few-shot classification accuracy. This semi-supervised setting can be extended to training stage as in (Ren et al., 2018), where they also consider the scenario in which the unlabeled support set may contain samples from irrelevant classes.

6.3 Experiments and Results

For fair comparison we follow the same experiment setting as in most recent few-shot learning works (Snell et al., 2017; Sung et al., 2018; Vinyals et al., 2016). We evaluate our approach on

three related tasks: few-shot classification on Omniglot (Lake et al., 2011) and miniImagenet (Vinyals et al., 2016), zero-shot classification on Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011), and semi-supervised few-shot adaptation on miniImagenet. Like most few-shot learning models we utilize four convolutional blocks for the embedding module to make the experiments comparable. Specifically, each convolutional block comprises a 64-filter 3×3 convolution, batch normalization layer (Ioffe and Szegedy, 2015), a ReLU nonlinearity, and a 2×2 max-pooling layer. When applied to the 28×28 Omniglot images this architecture results in a 64-dimensional embedding. We use the same encoder for embedding (i.e., f_ϕ) of both support and query points. The query embedding is modified further by a 1-D CNN in query encoder.

The 1-D CNN in the query embedding module has a convolutional block, batch normalization, and non-linear activation ReLU. As the input tensors are one-dimensional representations of the query images, the convolutional filters are one dimensional of size 1×3. Our model is trained end-to-end via SGD with Adam (Kingma and Ba, 2014). We use an initial learning rate of 10^{-3} and cut the learning rate in half every 2000 episodes. We observe that the classification performance of our model remains largely stable across a wide range of small λ values, so we fix it to 0.005. Also, the γ in Eq. 6.7 is set to 0.01. Optimizing the loss in our model does not need any complex selection of the training samples such as pairs or triplets. Consequently, the learning of our CNN based model is more efficient than methods based on contrastive or triplet loss and is easy to implement. Moreover, the learning objective of our loss is intra-class compactness, which is critical for discriminative feature learning in few-shot classification. We implement our model using the TensorFlow (Abadi et al., 2016) deep learning framework on an Intel Xeon CPU and two NVIDIA TITAN X GPU.

6.3.1 Few-Shot Learning

We perform experiments for few-shot classification on Omniglot (Lake et al., 2011) and miniImagenet (Vinyals et al., 2016) datasets as follows.

Table 6.2: Omniglot few-shot classification. Results are accuracies averaged over 1000 test episodes and with 95% confidence intervals where reported.

Model	Fine Tune	5-way Acc.		20-way Acc.	
		1-shot	5-shot	1-shot	5-shot
MANN (Santoro et al., 2016)	N	82.8%	94.9%	-	-
C SIANETS (Koch et al., 2015)	N	96.7%	98.4%	88.0%	96.5%
CL SIA NETS (Koch et al., 2015)	Y	97.3%	98.4%	88.1%	97.0%
MNETS (Vinyals et al., 2016)	N	98.1%	98.9%	93.8%	98.5%
M NETS (Vinyals et al., 2016)	Y	97.9%	98.7%	93.5%	98.7%
SIAMEMORY (Kaiser et al., 2017)	N	98.4%	99.6%	95.0%	98.6%
NSTAT(Edwards and Storkey, 2016)	N	98.1%	99.5%	93.2%	98.1%
METNETS (Munkhdalai and Yu, 2017)	N	99.0%	-	97.0%	-
MAML (Finn et al., 2017)	Y	98.7%	**99.9**%	95.8%	98.9%
RELATION NET (Sung et al., 2018)	N	99.6%	99.8%	97.6%	99.1%
PROTO NETS (Snell et al., 2017)	N	98.8%	99.7%	96.0%	98.9%
CDFS (ours)	N	**99.7**%	99.8%	**98.4**%	**99.5**%

Omniglot

Omniglot dataset contains 1623 characters (classes) from 50 different alphabets. There are 20 samples in each class, drawn by different people. For this experiment all input images are resized to 28×28. Following previous few-shot classification works, we augment new classes through 90, 180 and 270 rotations of existing data and use 1200 original classes plus rotations for training and remaining 423 classes plus rotations for testing. The few-shot classification accuracy on Omniglot is computed by averaging over 1000 randomly generated episodes from the testing set. During training, the 5-way 1-shot contains 19 query images, the 5-way 5-shot has 15 query images, the 20-way 1-shot has 10 query images and the 20-way 5-shot has 5 query images in each episode. The total number of samples in each episode for different settings during training is show in Table 6.1.

During testing, there are one and five query images per class for the 1-shot and 5-shot experiments, respectively. The results of 5-way and 20-way classification for 1-shot and 5-shot classification are shown in Table 6.2. The best-performing methods are highlighted. The proposed method achieves state-of-the-art performance under 20-way experiments setting and competitive results for 5-way classification. Specifically the proposed method has

Table 6.3: miniImageNet few-shot classification. Results are accuracies averaged over 600 test episodes and with 95% confidence intervals where reported.

Model	**Fine Tune**	**5-way Acc.**	
		1-shot	5-shot
MATCHING NETS (Vinyals et al., 2016)	N	43.5%	55.3%
META NETS (Munkhdalai and Yu, 2017)	N	49.2%	-
MAML (Finn et al., 2017)	Y	48.7 %	63.1%
META-LEARN LSTM (Ravi and Larochelle, 2016)	N	43.4%	63.1%
RELATION NET (Sung et al., 2018)	N	50.4%	65.3%
PROTOTYPICAL NETS (Snell et al., 2017)	N	49.4%	68.2%
CDFS (ours)	N	**52.7**%	**72.8**%

improved its baseline based on Prototypical Nets. (Snell et al., 2017). For 5-way 5-shot setting almost all methods perform perfectly since it is a rather easy classification task.

*mini*ImageNet

The *mini*Imagenet dataset, consists of $60,000$ RGB images with 100 classes, each having 600 examples and we resize input images to 84×84. 64, 16, and 20 classes are used for training, validation and testing, respectively. During training, there are 80 and 75 images in one episode of 5-way 1-shot and 5-way 5-shot setting. In fact, the 5-way 1-shot setting contains 15 query images, and 5-way 5-shot setting has 10 query images for each of the N_C classes in each training episode. Few-shot classification accuracies on *mini*Imagenet are shown in Table 6.3. All accuracy results are averaged over 600 test episodes and are reported with 95% confidence intervals. For each experiment setting the best-performing method is highlighted. The proposed method achieves state-of-the-art result in both 1-shot and 5-shot settings without any fine-tuning. Table 6.3 shows that our proposed method can improve the 1-shot and 5-shot accuracy of Prototypical Nets by around 3% for 1-shot and 4% for 5-shot in 20-way classification.

6.3.2 Zero-Shot Classification

We use the Caltech-UCSD Birds (CUB) 200-2011 dataset in order to evaluate our proposed method for zero-shot learning. The CUB dataset contains $11,788$ images of 200 bird species.

Table 6.4: Zero-shot classification accuracies on CUB-200.

Model	Feature Ext.	50-way Acc.
SJE (Akata et al., 2015)	GoogLeNet	50.1
ESZSL (Romera-Paredes and Torr, 2015)	GoogLeNet	47.2
SSE-RELU (Zhang and Saligrama, 2015)	VGG-19	30.4
JLSE (Zhang and Saligrama, 2016)	VGG-19	42.1
SYNC-STR(Changpinyo et al., 2016)	GoogLeNet	54.5
SEC-ML (Bucher et al., 2016)	VGG-19	43.3
REL. NET (Sung et al., 2018)	N-GoogLeNet	**62.0**
PROTO.NETS (Snell et al., 2017)	GoogLeNet	54.6
CDFS (ours)	GoogLeNet	55.8

We divide the classes into 100 training, 50 validation, and 50 test. For images we use 1024-D features extracted by applying GoogLeNet (Szegedy et al., 2015) pre-trained on ImageNet. We also augment images using the procedure in (Snell et al., 2017). For class attribute for zero-shot setting the 312-dimensional attribute vectors provided with the CUB dataset are used. These attributes encode various characteristics of the bird species such as their color, shape, and feather patterns.

We use an MLP network on top of both the 1024-dimensional image features and the 312-dimensional attribute vectors to produce a 1024-dimensional output space. We normalize the class prototypes to be of unit length, since the attribute vectors come from a different modality than the images. Training episodes were constructed with 50 classes and 10 query images per class. The embeddings were optimized via SGD with Adam at a fixed learning rate of 10^{-4}. The result of zero-shot learning is shown in Table 6.4. The second column demonstrate the type of feature extractor that these methods use for extracting image features (i.e., either VGG-19 (Sung et al., 2018) or GoogLeNet (Szegedy et al., 2015)).

It can be observed that Relation Net. (Sung et al., 2018) outperforms other methods. However, this method is not directly comparable with other methods since its image embedding subnet and how the visual feature space is computed are slightly different from other methods as discussed in (Sung et al., 2018). Our proposed method is the second best performing approach in zero-shot learning.

6.3.3 Semi-supervised Adaptation

We assume that we have access to a few labeled examples (i.e., five example per class) and many unlabeled examples from the same classes in the support set. Since our model is able to generate highly distinguishable feature embeddings in form of separate clusters, unlabeled samples are clustered to the corresponding classes in test time. The prototypes are estimated at test time using the labeled and unlabeled samples and then the query samples are classified based on the nearest prototype. We use *mini*Imagenet for this experiment and the 5-way 1-shot setting contains 15 query images, and 5-way 5-shot setting has 10 query images for each of the classes in each training episode. Table 6.5 shows how the number of unlabeled examples at test time affects the classification accuracy of the trained model. The results indicate that more unlabeled samples yield better performance, however, with the increase of the number of unlabeled samples, the improvement plateaus. Please note that semi-supervised setting is only for test time and the training needs the full label set.

6.3.4 Ablation Study

In order to evaluate the effect of context-aware query encoder and the $S3$ loss we perform the following ablation study. The first experiment setting is training and testing the model without using the query encoder which we denote by CDFS-NoQE. The second scenario is to remove the $S3$ loss during training which we denote by CDFS-NoS3. For this experiment the *mini*Imagenet is used in 5-way 5-shot setting and all the experimental parameters are the same as in Sec. 6.3.1. It can be observed from Table 6.6 that removing either the $S3$ regularization or the query encoder causes the performance to drop since it reduces the discriminative power of the model. However, removing the $S3$ loss has more negative effect

Table 6.5: 5-way testing accuracy using CDFS method on *mini*Imagenet for the semi-supervised scenario for different number of unlabeled samples per class (n).

n	1-shot	5-shot
5	52.9	73.0
10	54.0	74.3
20	55.2	74.8
40	58.9	75.1

Table 6.6: Ablation study to evaluate the effect of S3 loss and query encoder in the CDFS model on miniImagenet dataset.

Model	5-way 5-shot
CDFS-NoQE	70.6
CDFS-NoS3	69.1
CDFS (full)	**72.8**

on accuracy and causes a drop of 3.7% in accuracy which shows the importance of this regularization in performance of the model.

6.4 Conclusion

In this research, we introduced a simple but effective few-shot learning model which can produce highly discriminative embedding space with low intra-class variance. With removing the softmax loss and defining the episodes as one set without a query, the proposed approach can be considered as a few-shot clustering method which learns a deep non-linear metric in order to learn to cluster the data in few-shot setting using meta-learning. The future work is to extend the proposed framework to unsupervised few-shot classification by following the idea of *learning to cluster* proposed in this work.

Chapter 7

Conclusion and Future Works

In this dissertation, the importance of attention mechanism in recognition tasks in computer vision was studied by designing novel attention-based models and four scenarios were investigated that represent the most important aspects of attention mechanism.

An attention-based model was designed to reduce the visual features' dimensionality by selectively processing only a small subset of the data. We studied this aspect of the attention mechanism in a framework based on object recognition in distributed camera networks. However, the proposed model for feature selection is not limited to camera networks and can be used in any scenario where the goal is to extract the most informative part of the data.

Furthermore, an attention-based image retrieval system (i.e., person re-identification) was proposed which can learn to focus on the most discriminative regions of the person's image and process those regions with higher computation power using a deep convolutional neural network. Furthermore, we showed how visualizing the attention maps can make deep neural networks more interpretable. In other words, by visualizing the attention maps we can observe the regions of the input image where the neural network relies on, in order to make a decision. Future work can be adding a temporal attention to the proposed model to study the effect of spatio-temporal attention on video-based recognition tasks.

Moreover, in this dissertation a model for estimating the importance of the objects in a scene based on a given task was proposed. The proposed model estimates the importance of the road users that a driver (or an autonomous vehicle) should pay attention to in a driving scenario in order to have safe navigation. In this research, we investigated the effect of ego

car's intention and its context on estimating road users importance using only images taken from 3 cameras in front of the car. The proposed iCARE model estimates the important road users based on a 2-stage recognition framework, where the first stage generates important road user proposals using an importance-guided training scheme. In the second stage, model selectively picks the most important road user proposals by taking into account the location and intention context information. Our future work is to incorporate the intention of the road users into our model which plays an important role in determining which road user is important. Furthermore, incorporating other contextual information (e.g., depth, motion, etc.) can be an interesting line of future research for road user importance estimation.

Last but not least, an attention-based module and a new loss function were proposed in order to incorporate the context of the task into the feature representations of the samples and increasing the few-shot recognition accuracy. In this research, we introduced a simple but effective few-shot learning model which can produce highly discriminative embedding space with low intra-class variance. With defining the episodes as one set without a query, the proposed approach can be considered as a few-shot clustering method which learns a deep non-linear metric in order to learn to cluster the data in few-shot setting using meta-learning. The future work is to extend the proposed framework to unsupervised few-shot classification by following the idea of learning to cluster proposed in this work. The approach of learning to learn, or meta-learning, is a key stepping stone towards versatile models that can continually learn a wide variety of tasks throughout their lifetimes. Regarding the exciting power of meta-learning to deal with new tasks, we expect to see great surge of interest for applications and improvements of meta-learning based models in the near future.

To sum up, in this dissertation, we showed that attention can be multi-facet and studied the attention mechanism from the perspectives of feature selection, reducing computational cost, interpretable deep learning models, task-driven importance estimation, and context incorporation. Regarding the effectiveness of attention mechanism in different aspects of computer vision and deep learning, in future, we expect to see developments of numerous novel attention-based models and use of these models in various new tasks as well.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation 16)*, pages 265–283. 78

Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3908–3916. 2, 4, 15, 17, 24, 39

Akata, Z., Reed, S., Walter, D., Lee, H., and Schiele, B. (2015). Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936. 81

Asadinejad, A., Rahimpour, A., Tomsovic, K., Qi, H., and Chen, C.-f. (2018). Evaluation of residential customer elasticity for incentive based demand response programs. *Electric Power Systems Research*, 158:26–36. 14

Ba, J., Mnih, V., and Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*. 8, 9, 38

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. xi, 8, 9, 10, 38

Bai, S., Bai, X., and Tian, Q. (2017). Scalable person re-identification on supervised smoothed manifold. *arXiv preprint arXiv:1703.08359*. 47

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *ECCV 2006*. Springer. 13

Bedagkar-Gala, A. and Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286. xi, 3

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*. 56

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122. 29

Bro, R. and De Jong, S. (1997). A fast non-negativity-constrained least squares algorithm. *Journal of chemometrics*, 11(5):393–401. 30

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688. 17, 39

Bucher, M., Herbin, S., and Jurie, F. (2016). Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pages 730–746. Springer. 81

Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S., Grzeszczuk, R., and Girod, B. (2009). Chog: Compressed histogram of gradients a low bit-rate feature descriptor. In *CVPR 2009.* IEEE. 13

Chang, C.-I. and Heinz, D. C. (2000). Constrained subpixel target detection for remotely sensed imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 38(3):1144–1159. 30

Changpinyo, S., Chao, W.-L., Gong, B., and Sha, F. (2016). Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336. 81

Chartrand, R. and Staneva, V. (2008). Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(3):035020. 28

Chen, D., Yuan, Z., Chen, B., and Zheng, N. (2016). Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1268–1277. 47

Chen, X., Ma, H., Wan, J., Li, B., and Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In *IEEE CVPR*, volume 1, page 3. 20, 53

Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N. (2016a). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344. 4, 15, 17, 18, 39, 44

Cheng, J., Dong, L., and Lapata, M. (2016b). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*. 10

Chopra, S., Hadsell, R., LeCun, Y., et al. (2005). Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546. 22, 69, 70

Christoudias, C. M., Urtasun, R., and Darrell, T. (2008). Unsupervised feature selection via distributed coding for multi-view object recognition. In *Computer Vision and Pattern Recognition. CVPR 2008. IEEE Conference on*. IEEE. 2, 13, 14, 24, 25

Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2018). Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154. 54

Das, A., Chakraborty, A., and Roy-Chowdhury, A. K. (2014). Consistent re-identification in a camera network. In *European Conference on Computer Vision*, pages 330–345. Springer. 16

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR), 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE. 18

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634. 18

Dong, J., Karianakis, N., Davis, D., Hernandez, J., and Balzer (2015). Multi-view feature engineering and learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 14

Edwards, H. and Storkey, A. (2016). Towards a neural statistician. *arXiv preprint arXiv:1606.02185*. 79

et al., M. A. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 43

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611. 22, 69

Ferrari, V., Tuytelaars, T., and Van Gool, L. (CVPR-2004). Integrating multiple model views for object recognition. In *Computer Vision and Pattern Recognition. Proceedings of the IEEE Computer Society Conference on*. IEEE. 2, 24

Finley, T. and Joachims, T. (2008). Training structural svms when exact inference is intractable. In *Proceedings of the 25th international conference on Machine learning*, pages 304–311. ACM. 74

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org. 21, 67, 69, 79, 80

Halterman, R. and Bruch, M. (2010). Velodyne hdl-64e lidar for unmanned surface vehicle obstacle detection. In *Unmanned Systems Technology XII*, volume 7692, page 76920D. International Society for Optics and Photonics. 20, 53

Han, D. and Kim, J. (2015). Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*, pages 5016–5023. 14

He, R., Tan, T., Wang, L., and Zheng, W.-S. (2012). l 2, 1 regularized correntropy for robust feature selection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2504–2511. IEEE. 14

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*. 47

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*. 31

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*. 44, 58, 78

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025. 11

Janai, J., Güney, F., Behl, A., and Geiger, A. (2017). Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *arXiv preprint arXiv:1704.05519*. 21, 54

Jiang, Z., Lin, Z., and Davis, L. S. (TPAMI-2013). Label consistent k-svd: Learning a discriminative dictionary for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 27

Jiang, Z., Zhang, G., and Davis, L. S. (CVPR-2012). Submodular dictionary learning for sparse coding. In *Computer Vision and Pattern Recognition, 2012 IEEE Conference on*. IEEE. 27

Joachims, T., Finley, T., and Yu, C.-N. J. (2009). Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59. 22, 70

Kaiser, Ł., Nachum, O., Roy, A., and Bengio, S. (2017). Learning to remember rare events. *arXiv preprint arXiv:1703.03129*. 79

Kalayeh, M. M., Basaran, E., Gökmen, M., Kamasak, M. E., and Shah, M. (2018). Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071. 46, 47

Karanam, S., Li, Y., and Radke, R. J. (2015). Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4516–4524. 19

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732. 10

Kaviani Baghbaderani, R. and Qi, H. (2019). Incorporating spectral unmixing in satellite imagery semantic segmentation. In *IEEE International Conference on Image processing*. IEEE. 14

Kaviani Baghbaderani, R., Wang, F., stutts, C., Qu, Y., and Qi, H. (2019). Hybrid spectral unmixing in land-cover classification. In *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 14

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 44, 59, 78

Klaser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association. 19

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2. 79

Kong, S. and Wang, D. (2012). A dictionary learning approach for classification: separating the particularity and the commonality. In *Computer Vision–ECCV 2012*. Springer. 27

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE. 16, 39, 47

Kuen, J., Wang, Z., and Wang, G. (2016). Recurrent attentional networks for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3677. 20, 53

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86. 28

Lake, B., Salakhutdinov, R., Gross, J., and Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33. 78

Layne, R., Hospedales, T. M., Gong, S., and Mary, Q. (2012). Person re-identification by attributes. In *Bmvc*, volume 2, page 8. 17

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788. 14, 15

Lee, J. J. (2008). Libpmk: A pyramid match toolkit. 26

Li, C., Wang, Z., and Qi, H. (2018). Fast-converging conditional generative adversarial networks for image synthesis. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2132–2136. IEEE. 17

Li, C., Zhou, W., and Yuan, S. (2015). Iris recognition based on a novel variation of local binary pattern. *The Visual Computer*, 31(10):1419–1429. 14

Li, W. and Wang, X. (2013). Locally aligned feature transforms across views. pages 3594–3601. 39, 44

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159. 17, 18, 39, 44

Li, W., Zhu, X., and Gong, S. (2017). Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*. 47

Li, X., Zhao, L., Wei, L., Yang, M.-H., Wu, F., Zhuang, Y., Ling, H., and Wang, J. (2016). Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, 25(8):3919–3930. 20, 53

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., and Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3610–3617. 15, 16

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206. 15, 16, 17, 39

Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., and Li, S. Z. (2010). Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1301–1306. IEEE. 15, 16, 39

Lin, H. and Bilmes, J. A. (2012). Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*. 75

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer. 56

Lisanti, G., Masi, I., Bagdanov, A. D., and Del Bimbo, A. (TPAMI-2015). Person re-identification by iterative re-weighted sparse ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2, 24

Liu, H., Feng, J., Qi, M., Jiang, J., and Yan, S. (2016). End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*. 40

Liu, K., Ma, B., Zhang, W., and Huang, R. (2015). A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3810–3818. 18, 19

Liu, L., Rahimpour, A., Taalimi, A., and Qi, H. (2017). End-to-end binary representation learning via direct binary embedding. In *IEEE, Image processing, arXiv preprint arXiv:1703.04960*. 17

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., and Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3557. 16

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee. 13

McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 18, 19

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. (2017). Meta-learning with temporal convolutions. *arXiv preprint arXiv:1707.03141*, 2(7). 23, 70

Mitra, K., Veeraraghavan, A., Sankaranarayanan, A. C., and Baraniuk, R. G. (2014). Toward compressive camera networks. *Computer*. 25

Munkhdalai, T. and Yu, H. (2017). Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org. 21, 69, 79, 80

Naikal, N., Yang, A. Y., and Sastry, S. S. (2010). Towards an efficient distributed object recognition system in wireless smart camera networks. In *Information Fusion (FUSION),13th Conference on*. IEEE. 25, 31, 32

Naikal, N., Yang, A. Y., and Sastry, S. S. (2011a). Informative feature selection for object recognition via sparse pca. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 818–825. IEEE. 14, 32

Naikal, N., Yang, A. Y., and Sastry, S. S. (2011b). Informative feature selection for object recognition via sparse pca. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE. 25, 33

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814. 59

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *CVPR 2006*, volume 2, pages 2161–2168. IEEE. 26

Oh Song, H., Jegelka, S., Rathod, V., and Murphy, K. (2017). Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390. 22, 70, 75

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987. 15, 16, 39

Oreshkin, B., López, P. R., and Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 719–729. 68

Palazzi, A., Abati, D., Calderara, S., Solera, F., and Cucchiara, R. (2017). Predicting the driver's focus of attention: the dr (eye) ve project. *arXiv preprint arXiv:1705.03854*. 20, 53

Park, S. J., Kim, T. Y., Kang, S. M., and Koo, K. H. (2003). A novel signal processing technique for vehicle detection radar. In *Microwave Symposium Digest, 2003 IEEE MTT-S International*, volume 1, pages 607–610. IEEE. 20, 53

Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325. 15, 16, 39

Pei, W., Tax, D. M., and van der Maaten, L. (2016). Modeling time series similarity with siamese recurrent networks. *arXiv preprint arXiv:1603.04713*. 18

Pugeault, N. and Bowden, R. (2015). How much of driving is preattentive? *IEEE Transactions on Vehicular Technology*, 64(12):5424–5438. 20, 53

Pugelj, M. and Džeroski, S. (2011). Predicting structured outputs k-nearest neighbours method. In *International Conference on Discovery Science*, pages 262–276. Springer. 74

Qian, M. and Zhai, C. (2013). Robust unsupervised feature selection. In *IJCAI*. Citeseer. 14

Rahimpour, A., Liu, L., Taalimi, A., Song, Y., and Qi, H. (2017a). Person re-identification using visual attention. In *IEEE International Conference on Image processing*. IEEE. 15

Rahimpour, A., Qi, H., Fugate, D., and Kuruganti, T. (2017b). Non-intrusive energy disaggregation using non-negative matrix factorization with sum-to-k constraint. *IEEE Transactions on Power Systems*, 32(6):4430–4441. 14

Rahimpour, A., Qi, H., Kuruganti, T., and Fugate, D. (2015). Non-intrusive load monitoring of hvac components using signal unmixing. In *The IEEE Global Conference on Signal and Information Processing (Global SIP 2015), Orlando, FL, USA*. IEEE. 14

Rahimpour, A., Taalimi, A., Luo, J., and Qi, H. (2016). Distributed object recognition in smart camera networks. In *IEEE International Conference on Image Processing, Phoenix, Arizona, USA*. IEEE. 13

Rahimpour, A., Taalimi, A., and Qi, H. (2017c). Feature encoding in band-limited distributed surveillance systems. In *ICASSP 2017-IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE. 13

Ravi, S. and Larochelle, H. (2016). Optimization as a model for few-shot learning. 21, 67, 69, 80

Redondi, A. E., Baroffio, L., Cesana, M., and Tagliasacchi, M. (2015). Cooperative features extraction in visual sensor networks: a game-theoretic approach. In *2015-Proceedings of the 9th International Conference on Distributed Smart Camera*. ACM. 2, 24, 25

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. (2018). Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*. 67, 77

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99. 54

Rinner, B. and Wolf, W. (2008). An introduction to distributed smart cameras. *Proceedings of the IEEE*, 96(10):1565–1575. 13

Romano, S., Vinh, N. X., Bailey, J., and Verspoor, K. (2016). Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666. 75

Romera-Paredes, B. and Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161. 81

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. 21, 22, 67, 69, 79

Saxena, A. and Rose, K. (2010). On scalable distributed coding of correlated sources. *Signal Processing, IEEE Transactions on*. 14

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823. 22, 41, 69, 70

Sharma, S., Kiros, R., and Salakhutdinov, R. (2015). Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*. 39, 40

Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., and Wang, J. (2015). Person re-identification with correspondence structure learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3200–3208. 16

Sheu, S.-T., Wu, J.-S., Huang, C.-H., Cheng, Y.-C., Chen, L., et al. (2007). Ddas: Distance and direction awareness system for intelligent vehicles. *Journal of information science and engineering*, 23(3):709–722. 20, 53

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087. 21, 22, 67, 68, 69, 71, 72, 77, 79, 80, 81

Song, Y., Zhang, Z., Rahimpour, A., and Qi, H. (2016). Dictionary reduction: Automatic compact dictionary learning for classification. In *The 13th Asian Conference on Computer Vision (ACCV 2016)*. 27

Su, C., Li, J., Zhang, S., Xing, J., Gao, W., and Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. *ICCV*. 47

Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L. S., and Gao, W. (2015). Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3739–3747. 15, 16, 39

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208. 21, 22, 67, 69, 72, 77, 79, 80, 81

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112. 18

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12. 55

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. 81

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826. 44

Taalimi, A., Rahimpour, A., Capdevila, C., Zhang, Z., and Qi, H. (2016a). Robust coupling in space of sparse codes for multi-view recognition. In *IEEE International Conference on Image Processing*. IEEE. 13

Taalimi, A., Rahimpour, A., Liu, L., and Qi, H. (2017). Multi-view task-driven recognition in visual sensor networks. In *IEEE International Conference on Image processing, 2017*. 13

Taalimi, A., Shams, H., Rahimpour, A., Khorsandi, R., Wang, W., Guo, R., and Qi, H. (2016b). Multimodal weighted dictionary learning. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 173–179. IEEE. 27

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Attention, Perception, & Psychophysics*, 49(1):83–90. 9

Tschiatschek, S., Iyer, R. K., Wei, H., and Bilmes, J. A. (2014). Learning mixtures of submodular functions for image collection summarization. In *Advances in neural information processing systems*, pages 1413–1421. 75

Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM. 22, 70

Turcot, P. and Lowe, D. (2009). Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV workshop on emergent issues in large amounts of visual data (WS-LAVD)*, volume 4. 14, 25, 33

Underwood, G., Humphrey, K., and Van Loon, E. (2011). Decisions about objects in real-world scenes are influenced by visual saliency before and during their inspection. *Vision research*, 51(18):2031–2038. 20, 53

Varior, R. R., Haloi, M., and Wang, G. (2016a). Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer. 17, 18, 47

Varior, R. R., Shuai, B., Lu, J., Xu, D., and Wang, G. (2016b). A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer. 18, 47

Varior, R. R., Wang, G., Lu, J., and Liu, T. (2016c). Learning invariant color features for person re-identification. In *IEEE Transaction on Image processing*. IEEE. 4, 15, 16, 39

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. 9

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638. 21, 22, 23, 67, 69, 70, 72, 77, 78, 79, 80

Wah, C., Branson, S., Perona, P., and Belongie, S. (2011). Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*, pages 2524–2531. IEEE. 78

Wang, F., Zuo, W., Lin, L., Zhang, D., and Zhang, L. (2016). Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 15, 17, 18

Wang, T., Gong, S., Zhu, X., and Wang, S. (2014). Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer. 18, 19

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256. 8, 9, 38

Xiao, T., Li, H., Ouyang, W., and Wang, X. (2016). Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 17, 18, 47

Xiong, F., Gou, M., Camps, O., and Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, pages 1–16. Springer. 15, 16

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81. xi, 8, 10, 11, 12, 38

Yang, A. Y., Maji, S., Christoudias, C. M., Darrell, T., Malik, J., and Sastry, S. S. (2010). Multiple-view object recognition in band-limited distributed camera networks. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–8. IEEE. 14

Yang, L. and Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2). 15

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., and Li, S. Z. (2014). Salient color names for person re-identification. In *European conference on computer vision*, pages 536–551. Springer. 16

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A. (2015). Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515. 11

Ye, Y., Ci, S., Katsaggelos, A. K., Liu, Y., and Qian, Y. (2013). Wireless video surveillance: A survey. *Access, IEEE*, 1:646–660. 13

Yeo, C., Ahammad, P., and Ramchandran, K. (2008). Rate-efficient visual correspondences using random projections. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE. 14

Yi, D., Lei, Z., Liao, S., Li, S. Z., et al. (2014). Deep metric learning for person re-identification. In *ICPR*, volume 2014, pages 34–39. 17

Yu, H., Wang, J., Huang, Z., Yang, Y., and Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593. 11

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702. 18

Zhang, L., Xiang, T., and Gong, S. (2016). Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4, 15, 17, 47

Zhang, Z. and Saligrama, V. (2015). Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174. 81

Zhang, Z. and Saligrama, V. (2016). Zero-shot learning via joint latent similarity embedding. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042. 81

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., and Tang, X. (2017a). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085. 47

Zhao, L., Li, X., Zhuang, Y., and Wang, J. (2017b). Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3219–3228. 47

Zhao, R., Ouyang, W., and Wang, X. (2013a). Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535. 15, 16, 39

Zhao, R., Ouyang, W., and Wang, X. (2013b). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3593. 15, 39

Zhao, R., Ouyang, W., and Wang, X. (2014). Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151. 16

Zheng, J., Jiang, Z., and Chellappa, R. (TIP-2016). Cross-view action recognition via transferable dictionary learning. *IEEE Transactions on Image Processing*. 24

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124. 15, 16, 19, 39, 44

Zhou, S., Wang, J., Wang, J., Gong, Y., and Zheng, N. (2017). Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 47

Zhu, P., Zuo, W., Zhang, L., Hu, Q., and Shiu, S. C. (2015). Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446. 14

Appendices

A Publications

- Context Aware Road User Importance Estimation For Autonomous Driving.
  A. Rahimpour, S. Martin, A. Tawari, H. Qi. *IEEE Intelligent Vehicles (IV2019).*

- Person Re-identification using visual attention.
  A. Rahimpour, H. Qi. IEEE International Conference on Image Processing 2017.

- Feature Encoding in Band-limited Distributed Surveillance Systems

  A. Rahimpour, A. Taalimi, H. Qi. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017), New Orleans, USA, 2017 - Selected as finalist for ICASSP 2017s Student Paper Contest.

- Non-Intrusive Energy Disaggregation Using Non-negative Matrix Factorization with Sum-to-k Constraint.

  A. Rahimpour, H. Qi, D. Fugate, T. Kuruganti. IEEE Transactions on Power Systems, 2017

- Distributed object recognition in smart camera networks.
  A. Rahimpour, A. Taalimi, J. Luo, H. Qi. IEEE International Conference on Image Processing (ICIP), Phoenix, Arizona, USA, 2016

- Non-Intrusive Load Monitoring of HVAC Components using Signal Unmixing.
  A. Rahimpour, H. Qi, D. Fugate, T. Kuruganti. The IEEE Global Conference on Signal and Information Processing (Global SIP), 2015.

- Evaluation of residential customer elasticity for incentive based demand response programs.
  A. Asadinejad, A.Rahimpour, K. Tomsovic, H. Qi. Electric Power Systems Research Journal, 2018

- Dictionary Reduction: Automatic Compact Dictionary Learning for Classification.
  Y. Song, Z. Zhang, L. Liu, A. Rahimpour, H. Qi. The 13th Asian Conference on Computer Vision (ACCV 2016)

- Robust coupling in space of sparse codes for multi-view recognition. A. Taalimi, A. Rahimpour, H. Qi. IEEE International Conference on Image Processing (ICIP), 2016

- Multimodal Weighted Dictionary Learning. A. Taalimi, A. Rahimpour, H. Qi. The IEEE Advanced Video and Signal-based Surveillance Conference (AVSS), 2016

- End-to-end Binary Representation Learning via Direct Binary Embedding. L. Liu, A. Rahimpour, H. Qi. IEEE International Conference on Image processing, 2017

- Multi-view Task-driven Recognition in Visual Sensor Networks. A.Taalimi, A. Rahimpour, L.Liu, H. Qi. IEEE International Conference on Image processing, 2017

- Eye Tracking by Image Processing for Helping Disabled People. A. Rahimpour, A. Nasiraei Moghaddam. Iranian Journal of Biomedical Engineering (IJBME) 6 (3), 195-205

- An Adaptive Template Matching Approach for Real time Eye Tracking and Facial Feature Extraction. A. Rahimpour, A. Nasiraei Moghaddam. Iranian Conference on Biomedical Engineering, ICBME 2012.

- Multivariate empirical mode decomposition based signal analysis and efficient-storage in smart grid. L. Liu, A. Albright, A. Rahimpour, J. Guo, H. Qi, Y. Liu. The IEEE Global Conference on Signal and Information Processing, 2016

Vita

Alireza Rahimpour received his B.S. degree in Electrical Engineering, Electronics from Shiraz University, Shiraz, Iran, in 2009 and the M.S. degree from Amirkabir University of Technology (Tehran Polytechnic) Tehran, Iran in 2012. He joined the Department of Biomedical Engineering, Azad University Tehran, Iran as a Lecturer in 2012. He has been with the Electrical Engineering and Computer Science Department at University of Tennessee at Knoxville, TN, USA as PhD candidate since 2013. He received the Electrical Engineering and Computer Science Department Excellence Fellowship and Outstanding Graduate Teaching Assistant Award from the University of Tennessee, Knoxville in 2013 and 2016 respectively. He also received the University of Tennessee Chancellors award and Outstanding Graduate Research Assistant award in 2017 and 2018. His current research interests include Machine learning, Computer Vision, Artificial intelligence, Signal and Image processing and Non-Intrusive Load Monitoring (NILM). Since 2013, he also has been collaborating as a researcher and mentor with Center for Ultra-Wide-Area Resilient Electric Energy Transmission Networks (CURENT), that is a National Science Foundation Engineering Research Center jointly supported by NSF (National Science Foundation) and the DoE (Department of Energy).